System-Wide Studies of Gene Expression in Escherichia coli by Fluorescence Microscopy and High Throughput Sequencing

Advisor: X. Sunney Xie Huiyi Chen

System-Wide Studies of Gene Expression in *Escherichia coli* by Fluorescence Microscopy and High Throughput Sequencing

Abstract

Gene expression is a fundamental process in the cell and is made up of two parts – the information flow from DNA to RNA, and from RNA to protein. Here, we examined specific sub-processes in *Escherichia coli* gene expression using newly available tools that permit genome-wide analysis. We begin our studies measuring mRNA and protein abundances in single cells by single-molecule fluorescence microscopy, and then focus our attention to studying RNA generation and degradation by high throughput sequencing.

The details of the dynamics of gene expression can be observed from fluctuations in mRNA and protein copy numbers in a cell over time, or the variations in copy numbers in an isogenic cell population. We constructed a yellow fluorescent fusion protein library in *E. coli* and measured protein and mRNA abundances in single cells. At below ten proteins per cell, a simple model of gene expression is sufficient to explain the observed distributions. At higher expression levels, the distributions are dominated by extrinsic noise, which is the systematic heterogeneity between cells.

Unlike proteins which can be stable over many hours, mRNA is made and degraded on the order of minutes in *E. coli*. To measure the dynamics of RNA generation and degradation, we developed a protocol using high throughput sequencing to measure steady-state RNA abundances, RNA polymerase elongation rates and RNA degradation rates simultaneously with high nucleotide-resolution genome-wide. Our data shows that RNA has similar lifetime at all positions throughout the length of the transcript. We also

find that our polymerase elongation rates measured *in vivo* on a chromosome are generally slower than rates measured on plasmids by other groups. Studying nascent RNA will allow further understanding of RNA generation and degradation. To this end, we have developed a labeling protocol with a nucleoside analog that is compatible with high throughput sequencing.

Acknowledgements

I don't think I fully realized what a huge commitment graduate school was going to be back when I decided to pursue a Ph.D. But I survived, and I learned to ask questions and do research, and continue to realize how little I know. I credit my professional and personal growth, as well as general wellbeing, to the colleagues, friends and family who have encouraged and mentored me. They have kept me passionate and curious about science and life.

When I picked Professor Sunney Xie as my advisor, I was expecting to be trained in research. Sunney has taught me that and more. He has shown me what it means to be a scientist and a citizen of the world. I am thankful for his patience and guidance through the slow times.

I would also like to thank my thesis committee, Dr. Richard Losick, Dr. Andrew Murray, and Dr. Eric Rubin. They have kept me on track these past 6.5 years, and have been invaluable for providing alternative points of view.

During my time in graduate school, I have had the privilege of learning from many individuals through work on various projects. My initial projects were on coherent anti-Stokes Raman scattering, and during that period I worked with Dr. Weiyuan Yang, and later Dr. Brian Saar. I worked on the YFP library project (chapter 2) with Dr. Paul Choi, and Dr. Gene-Wei Li, and Dr. Yuichi Taniguchi. We benefited from discussion with Dr. Uri Alon (Weizmann Institute) and Dr. Johan Paulsson (Harvard Medical School). For the RNA project (chapter 3) and metabolic labeling of RNA (chapter 4), I worked extensively with Dr. Katsuyuki Shiroguchi, who has been a great partner and mentor. Alec Chapman helped us set up data processing on the Odyssey and lab clusters. We benefited from discussion with Dr. Peter Sims, Professor Hao Ge (Peking University), Dr. Paul Choi, Dr. Will Greenleaf, Dr. Gene-Wei Li, Dr. Yuichi Taniguchi, Dr. Chenghang Zong, Dr. Steve Mao, and Dr. Rahul Roy. Lastly for the transposon-sequencing experiment, I worked with Dr. Guoqing Zhang, Dr. Xiaohui Ni, Professor Hao Ge, and Dr.

Haifeng Duan. We benefited from technical advice from Dr. Tim van Opijnen (Tufts Medical School), and discussion with Jason Zhang (Harvard School of Public Health).

I have also learned in many ways from my other colleagues, Dr. Nam Ki Lee, Dr. Giuseppe Lia, Dr. Martin Vogel, Dr. Sangjin Kim, Srinjan Basu, Dr. Erik Brostromer, and Dr. Christof Gebhardt. Ms. Teri Howard has been a constant source of encouragement, and recently Ms. Tracey Schaal has been important for keeping group morale high. I would also like to acknowledge in particular Dr. Peter Sims for looking out for me during the stressful last months of the Ph.D.

I would also like to thank the other staff at Harvard University for help with research and also for making graduate school more pleasant. They include Christian Daly, Brian Tilton, Jennifer Couget (FAS Center for Systems Biology), Jiangwen Zhang (Research Computing), Joe and Stan (VWR Stockroom), Emmanuel and other security guards whose names I don't know. I would also like to thank the MCB administration, in particular Michael Lawrence for helping me keep the paperwork straight.

My studies at Harvard may not have happened without help from Dr. Su Guaning, Dr. James Tam, and Dr. Alex Law (Nanyang Technological University).

Most of all, I would like to thank my family for being patient and understanding that I am never at home, and for their whole-hearted belief in me.

Table of Contents

List of Figures and Tables

Chapter 1: Introduction

1.1 Aim of this work

Gene expression is the process by which information stored in DNA is retrieved and made into a functional gene product, usually a protein, to alter the cell. While the flow of information is neatly summarized by the central dogma, which shows information flowing from DNA to RNA to protein, the regulation of this information flow is complex.

There are many modes of regulation at each step of gene expression. We look at the example of transcriptional regulation – one or more transcription factors control initiation of transcription for a particular gene. The methylation state of DNA helps determine whether the gene is transcribed. Other protein factors regulate elongation and termination. Furthermore, the RNA transcript could have small molecule sensing abilities, and form structures to cause RNA polymerase to prematurely abort transcription [18]. There are similarly a myriad of factors to consider at the downstream steps of splicing (where applicable), translation, and even post-translational modification. There are thus many questions to address when studying gene expression.

Because some proteins work with a variety of substrates, asking questions about these proteins is sometimes really asking a question about a system. Such proteins include RNA polymerase, degradation enzymes, the spliceosome, and the nucleosome. In many cases, a handful of specific genes or templates serve as the point of reference for a field. Without knowledge about the context, it is hard to know if the genes being studied are typical or outliers. System-wide datasets are thus important as a reference for these

few measurements. They allow us to assess the relative contribution of different factors and to identify outliers, which may indicate new modes of regulation.

In this work, we describe the development and use of resources and methods for generating system-wide information in *Escherichia coli* to address questions that naturally span an entire system. In Chapter 2, we report the construction of a chromosomal fluorescent protein fusion library which we used to survey noise, the cell-to-cell variation, in protein and mRNA. In Chapter 3, we describe the measurement of the dynamics of RNA synthesis and degradation genome-wide in *E. coli*, and propose a correction to the conventional protocol for measuring RNA degradation rate [2, 30]. Chapter 4 documents the development of a metabolic labeling scheme that is compatible with high throughput sequencing to study nascent RNA in *E. coli*. In the Appendix, we report preliminary efforts to construct a network of genes involved in bacterial persistence using transposons and high throughput sequencing.

1.2 Technology review

The availability of tools that facilitate parallel or high throughput data collection is a major contribution of the shift towards system-scale experiments in biology. We briefly review the tools available for studying nucleic acids and proteins and their applications to system-wide studies in the rest of this chapter. These tools are used to make different types of measurements. For instance, some methods can detect signals in single cells, which can address questions about the uniformity of gene expression in a population, while other methods are suitable for following dynamics in live cells.

1.2.1 Microarrays

Nucleic acids are easy to amplify and identify by their sequence. One of the most common systems-wide experiments with nucleic acids is gene expression profiling by microarray. This is typically a population-average, single-point-in-time type of experiment. DNA oligomers, designed to hybridize with specific sequences, are printed onto a glass microscope slide [28]. Fluorescent complementary DNA (cDNA) molecules are generated from the mRNA pool of interest by reverse transcription, and hybridized onto the array on the slide. The level of gene expression is measured by fluorescence intensity. Gene expression profiling has revealed the differences and similarities across cell types [16] and tissues [6], and can identify genes of interest for further studies.

Microarrays were a key tool in the first draft of the complete human transcriptome, leading to the discovery of many previously unknown transcripts [3]. Other interesting applications of microarrays include measuring kinetics by taking multiple points, like RNA degradation rates [2], and also probing nucleosome position and mobility across the genome [37].

1.2.2 High throughput sequencing

High throughput sequencing is a newer technology than microarrays that can serve a similar purpose. High throughput sequencing platforms were initially designed to sequence genomic DNA by the shotgun approach (as an alternative to conventional Sanger sequencing) – millions of short reads, initially around 30 bases but the read length has since been extended to up to 150 bases (or 800 bases if using 454 technology), are generated in parallel and pieced together by computer algorithms to produce a longer read. To extend the technology to RNA, a reverse transcription step is added to convert RNA to cDNA. The later steps of library construction, adding adapters that allow the

DNA molecules to amplify and attach to a surface, are the same. Further modifications can be made to the sample preparation to preserve the directionality of the RNA read (not relevant for DNA). RNA sequencing experiments are usually referred to as RNA-seq.

Several commercial sequencing platforms are currently in use, of which the main ones are 454 [22], Illumina [1], and SOLiD [31]. These technologies use sequencing, either by a polymerase or a ligase, for detection. Each read is mapped to a gene, and the total number of reads for each gene is counted. Unlike hybridization which suffers from non-specific hybridization, there is little chance of non-specific detection by sequencing. This makes high throughput sequencing a more quantitative method of measuring nucleic acids than microarray.

Detection by sequencing confers high throughput sequencing important advantages over microarrays. No prior knowledge of the genome or transciptome is needed – it is thus possible to detect unexpected RNAs, splicing products or alleles. Experiments are not limited by the range of available commercial microarrays, usually a problem when working with less common organisms. Moreover, the probe hybridization background is no longer a consideration, greatly simplifying experimental design. High throughput sequencing offers basepair resolution, although tiling arrays can be made to go down to 5bp resolution [3]. Lastly, high throughput sequencing offers a higher dynamic range of detection than microarray.

Several notable applications of high throughput sequencing to systems include the observation of translating ribosomes [15], and transcribing polymerases [8] by the Weissman lab. The ribosome study was able to identify distinct phases in translation, while the RNA polymerase study was able to observe large amounts of antisense

transcription across the yeast genome, and identify pause sequences. These applications benefit greatly from the basepair resolution of the technique. The 3D architecture of genomes can be interrogated by Hi-C, a sequencing-based method used to map physical proximity of regions of the genome to understand genome architecture [19]. Another contribution made with high throughput sequencing is the discovery of long intergenic non-coding RNAs (lincRNA) in mammals [13], although microarrays could probably have been used instead.

In certain cases, it may be informative to repeat experiments previously performed with microarray using high throughput sequencing. In this work, we discuss new information obtained from repeating the microarray measurement of RNA degradation by high throughput sequencing (Chapter 3).

1.2.3 RNA fluorescence in situ hybridization

While most microarray and high throughput sequencing experiments are used to study a population, these technologies can be applied to single cells [32, 33]. These single cell experiments can provide information about all genes in a cell. However, they are difficult and expensive to realize, thus there are few other single cell datasets for comparison.

A motivation for doing single cell experiments is that the population average is not always reflective of all cells. For instance, a bulk measurement of lacZ abundance will not detect the two subpopulations with different lacZ expression at intermediate inducer conditions [7, 25]. Population measurements cannot provide information about the spread of the distribution

Counting of RNA by fluorescence in situ hybridization (FISH) can give information about the quantity and spatial location of the RNA [10, 26]. The technique was further extended to single molecule FISH to confidently count the number of transcripts for a gene in single cells across a population [27]. To detect RNA transcripts, many fluorescent probes are designed to hybridize with specific parts of the gene [10, 27]. Alternatively, many RNA repeats are added on to the 3' end of a gene, and a single fluorescent oligomer probe is needed to detect the tandem repeats [26].

1.2.4 System-wide approaches used to detect proteins

A different set of tools is used to study proteins. These include forms of mass spectrometry, 2D gels and western blots. In general, mass spectrometry and 2D gels have been useful for studying the ~500 most abundant proteins in the cell [20, 21]. Western blot, a more sensitive technique, was used to quantify the entire range of proteins in yeast strains [12]. The availability of antibodies limits the number of native proteins that can be studied. In this case, each gene was tagged with a tandem affinity purification tag so that one antibody can be used against all genes. Besides short peptide tags, another useful protein tag is green fluorescent protein (GFP).

1.2.5 Fluorescent proteins

GFP is a jellyfish-derived protein with a native chromophore. Since the first reported cloning and expression of GFP in *Escherichia coli* and *Caenorhabditis elegans* [5], GFP and its variants have been used in a wide variety of gene expression experiments ranging in scale from single molecule imaging [7, 9, 36] to whole body imaging [35]. Unlike purification tags that are only good for single time point experiments, GFP can be

tracked over time to detect dynamics in living cells and organisms. This distinguishes GFP from other protein detection methods.

While fluorescent proteins (FP) have been used to tag single genes for specific experiments, the general usefulness of FPs have prompted the construction of libraries in organisms. These libraries have been disseminated to different laboratories worldwide. FPs have been used as promoter reporters [38], fused to genes on plasmids [17] and fused to genes in native chromosomal positions in the resource libraries that have been created in bacteria [23], yeast [14] and mammalian cells [29]. The libraries were used to survey protein abundance, noise and localization genome-wide, and were useful in discovering uncharacterized transcription units and temporal patterns of localization. The availability of these resources facilitates other system-wide studies; the yeast GFP library was used to uncover widespread frequency modulation in gene expression [4]. In the above studies, FPs are primarily detected by flow cytometry or fluorescence microscopy.

1.2.6 Flow cytometry

In flow cytometry, cells are suspended in liquid stream that is broken into droplets, each containing one cell. These cells are interrogated by a laser beam, and the information, such as wavelength, fluorescence intensity and scattering, is collected. Flow cytometers were first designed to sort by volume and had the capacity to handle 500 – 1000 cells per second [11]. Current flow cytometers can handle multiple fluorescent signals and process up to 70,000 events per second. Flow cytometry is thus highly suited to high throughput data collection. It was used to measure protein abundance in thousand

of single yeast cells for more than 4000 strains [24]. However, flow cytometry has limited sensitivity, so less abundant yeast proteins could not be detected in the study.

1.2.7 Fluorescence microscopy

Microscopy is more sensitive than flow cytometry while still capable of measuring large signals, and can provide additional information about the localization of the tagged product. It was thus the tool of choice for the system-wide protein localization study in yeast [14]. The throughput of a microscope is lower than flow cytometry (<70,000 cells/second), but the efficiency of data collection on a basic microscope setup can be improved by use of automated software and parts like a stage scanner.

Fluorescence microscopy is capable of detecting down to a single FP given the right conditions [34]. It is important to keep the cellular autofluorescence low by growing cells in the right medium, for instance M9 medium for *E. coli*. Using yellow FP improves the signal collection as the excitation and emission peaks are away from the autofluorescence peaks. It is also helpful to localize the fluorescence signal by immobilizing the FP by fusing it with a membrane protein [7, 36], or a DNA-binding protein [9] to allow direct observation of the fluorescence signal. Of course, a sensitive camera, usually an EM-CCD, should be used for imaging.

Single molecule imaging is particularly important in bacterial cells, which have the same concentration of proteins as the larger yeast cells but are only $1/50^{th}$ of the volume. DNA and RNA molecules are also present in small numbers in the bacterial cell.

Another important feature of fluorescence microscopy is that it is non-invasive, allowing researchers to track dynamics in single live cells. It would be impossible to

watch protein expression in near real-time [7, 36], or to measure the frequency of transcription factors shuttling between the nucleus and cytoplasm [4] with another technique.

1.3 Summary

We reviewed and highlighted several advances in technology that have permitted the systems-wide studies that have become increasingly common in biological research. We discuss new applications of these techniques to address questions in gene expression in *E. coli* in the following chapters.

1.4 References

1. Bentley, D.R., et al. (2008) "Accurate whole human genome sequencing using reversible terminator chemistry." *Nature* **456**, 53 59

2. Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S., Cohen, S.N. (2002) "Global Analysis of mRNA Decay and Abundance in *Escherichia coli* at Single-Gene Resolution Using Two-Color Fluorescent DNA Microarrays." *Proc. Nat. Ac. Sci.* **99**, 9697 – 9702

3. Bertone, P. et al. (2004) "Global Identification of Human Transcribed Sequences with Genome Tiling Array." *Science* **306**, 2242 – 2246

4. Cai, L., Dalal, C.K., Elowitz, M.B. (2008) "Frequency-modulated nuclear localization bursts coordinate gene expression." *Nature* **455**, 485 – 490

5. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., Prasher, D.C. (1994) "Green Fluorescent Protein as a Marker for Gene Expression." *Science* **263**, 802 – 805

6. Chan, E.T. et al. (2009) "Conservation of Core Gene Expression in Vertebrate Tissues." *J. Biol.* **8**, 33.

7. Choi, P.J., Cai, L., Frieda, K., Xie, X.S. (2008) "A Stochastic Single Molecule Event Triggers Phenotype switching in a bacterial cell." *Science* **322**, 442 – 446

8. Churchman, L.S., Weissman, J.S. (2010) "Nascent transcript sequencing visualizes transcription at nucleotide resolution." *Nature* **469**, 368 – 373

9. Elf, J., Li, G.-W., Xie, X.S. (2007) “Probing Trancription factor dynamics at the single molecule level in a single cell.” *Science* **316**, 1191 – 1194

10. Femino, A.M., Fay, F.S., Fogarty, K. Singer, R.H. (1998) “Visualization of single RNA Transcripts in situ” *Science* **280**, 585 – 590

11. Fulwyler, M.J. (1965) “Electronic Separation of Biological Cells by Volume.” *Science* **150**, 910 – 911

12. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O’Shea, E.K., Weissman, J.S. (2003) “Global Analysis of Protein Expression in Yeast.” *Nature* **425**, 737 – 740

13. Guttman, M. et al. (2009) “Chromatin Signature Reveals over a thousand highly conserved large non-coding RNAs in mammals.” *Nature* **458**, 223 – 227.

14. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O’Shea, E.K. (2003) “Global Analysis of Protein Localization in Budding Yeast.” *Nature* **425**, 686 – 691

15. Ingolia, N.T., Ghaemmagami, S., Newman, J.R.S., Weissman, J.S. (2009) “Genome-wide Analysis in vivo of translation with nucleotide resolution using ribosome profiling.” *Science* **324**, 218 – 223

16. Kai, T., Williams, D., Spradling, A.C. (2005) “The Expression Profile of Purified Drosophila germline stem cells.” *Dev. Bio.* **283**, 486 – 502

17. Kitagawa, M., Ara, T., Arifuzzman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., Mori, H. (2005) “Complete set of ORF clones of *Escherichia coli* ASKA library (A complete set of *E. coli K-12* ORF Archive): Unique Resources for Biological Research.” *DNA Res.* **12**, 291 – 299

18. Landick, R., Turnbough, C.L.Jr., Yanofsky, C. (1996) “*Escherichia coli and Salmonella: Transcription Attenuation”* Edited by Neidhardt, F.C., 1265 – 1268

19. Lieberman-Aiden, E., van Berkum, N.L. et al. (2009) “Comprehensive Mapping of Long Ranged Interactions Reveals Folding Principles of the Human Genome.” *Science* **326**, 289 – 293

20. Lopez-Campistrous, A. et al. (2005) “Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth.” *Mol. Cell. Proteomics* **4**, 1205 – 1209

21. Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M. (2005) “Abosolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.” *Nat. Biotech.* **25**, 117 – 124

22. Margulies, M., Egholm, E. et al (2005) “Genome sequencing in microfabricated high-density picolitre reactors.” *Nature* **437**, 376 – 380

23. Meile, J.-C., Wu, L.J., Ehrlich, S.D., Errington, J., Noirot, P. (2006) “Systematic localization of proteins fused to the green fluorescent protein in Bacillus subtilis: Identification of New Proteins at the DNA Replication Factory.” *Proteomics* **6**, 2135 – 2146.

24. Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., Weissman, J.S. (2006) “Single-cell Proteomic Analysis of *S. cerevisiae* Reveals the Architecture of Biological Noise.” *Nature* **441**, 840 – 846

25. Novick, A., Weiner, M. (1957) “Enzyme Induction As An All or None Phenomenon.” *Proc. Nat. Ac. Sci.***, 43**, 553 – 566

26. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., Tyagi, S. (2006) “Stochastic mRNA expression in mammalian cells.” *PLoS Biol.* **4(10)**, e309

27. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., Tyaji, S. (2008) “Imaging individual mRNA molecules using multiply single labeled probes.” *Nat. Methods* **5**, 877 – 879

28. Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.” *Science* **270**, 467 – 470

29. Segal, A., et al. (2006) “Dynamic Proteomics in Individual Human Cells Uncovers Widespread Cell-Cycle Dependence of Nuclear Proteins.” *Nat. Methods* **3**, 525 – 531

30. Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., Rosenow, C. (2003) “Global RNA Half-Life Analysis in *Escherichia coli* Reveals Positional Patterns of Transcript Degradation” *Genome Research* **13**, 216 -223

31. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M. (2005) “Accurate multiplex polony sequencing of an evolved bacterial genome.” *Science* **309**, 1728 – 1732

32. Tang, F. et al. (2009) “mRNA-seq Whole-Transcriptome Analysis of a Single Cell.” *Nat. methods*, **6**, 377 – 382

33. Tietjen, I., Rihel, J.M., Cao, Y., Koentges, G., Zakhary, L., Dulac, C. (2003) "Single-Cell Transcriptional Analysis of Neuronal Progenitors." *Neuron,* **38**, 161 -175

34. Xie, X.S., Choi, P.J., Li, G.-W., Lee, N.K., Lia, G. (2008) "Single Molecule Approach to Molecular Biology in Living Cells." *Annu. Rev. Biophys.* **37**, 417 – 444

35. Yang, M., et al. (2000) "Whole-body Optical Imaging of Green Fluorescent Protein-expressing tumors and metastases." *Proc. Nat. Ac. Sci.* **97**, 1206 – 1211

36. Yu, J, Xiao, J., Ren, X, Lao, K., Xie, X.S. (2006) "Probing Gene Expression in Live Cells, One Protein at a Time." *Science* **311**, 1600 – 1603

37. Yuan, G.-C., Liu, Y.-J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., Rando, O.L. (2005) "Genome-scale Identification of Nucleosome Positions in *S. cerevisiae.*" *Science* **309**, 626 – 630.

38. Zaslaver, A., et al. (2006) "A Comprehensive Library of Fluorescent Transcriptional Reporters for *Escherichia coli.*" *Nat. Methods* **3**, 623 – 628

Chapter 2: Survey of abundance and noise of protein and RNA genome-wide in Escherichia coli

Contributions:

Prof. Sunney Xie initiated the development of an *E. coli* chromosomal fluorescent reporter library. This project was a collaboration with Dr. Andrew Emili (University of Toronto), who had a library of tagged *E. coli* genes. Dr. Mohan Babu sent the *E. coli* strains from Canada. Dr. Choi designed the library, and he and I constructed the YFP library, with assistance from Mr. Jeremy Hearn. I verified the strains by sequencing, and prepared samples for use in the microfluidic devices designed by Dr. Yuichi Taniguchi. Dr. Taniguchi performed the steady state protein imaging and analysis. High-expression real time experiments, and two-color protein imaging experiments were performed by Dr. Taniguchi and Dr. Choi, while I performed the single-molecule real-time experiments. Dr. Gene-Wei Li conceived the idea of applying RNA FISH to the reporter library and first implemented it. Working with Dr. Taniguchi, he collected the final data. I measured genome-wide RNA abundances and lifetimes using RNA-seq, with Dr. Choi providing assistance with data analysis. Dr. Choi, Dr. Taniguchi, and Dr. Li contributed to the modeling. This work was published in *Science* **329**, 533 (2010).

2.1 Abstract

Protein and messenger RNA (mRNA) exist in low copy numbers, and their numbers vary from cell to cell in isogenic bacterial populations. Using a newly-created yellow fluorescent protein fusion library in *Escherichia* coli, we carried out system-wide measurement of protein and mRNA abundance in single cells with single molecule

sensitivity. We obtained the population distribution (noise) of each protein and highly abundant mRNA. At below 10 proteins per cell, distributions are dominated by intrinsic noise (biochemical fluctuations) and at higher expression levels, they are dominated by extrinsic noise (differences between cells). The cell minimizes noise by expressing most essential genes at above 10 copies per cell. mRNA and protein abundances are not correlated in single cells, an effect of the large differences in half-lives, and also possibly noise in translation.

2.2 Introduction

The level of gene expression is a common and informative readout of an experiment or laboratory procedure. It reflects both the history and the future of the cell – what the cell is responding to, and what its response is. Gene expression is often measured as an average across a population, masking the variability of gene expression from cell to cell, or noise. Noise is the variation in gene expression between cells of an isogenic population, or as the fluctuations across time of a cell lineage. Both manifestations of noise are linked [11]. Noise may be the basis for helping cells cope and survive ever-changing environments [7, 16].

Protein noise was initially studied in yeast and bacteria cells on a limited basis [9, 22, 25]. The first system-wide surveys of noise were performed in yeast [1, 21]. Briefly, proteins were tagged with a green fluorescent protein were expressed from their native promoters on the chromosome, and the amount of protein in single cells was measured by flow cytometry. However, flow cytometry was unable to distinguish the levels of less abundant protein from the cell's autofluorescence background. Thus a little over 1/3 of the least abundant proteins in yeast were not studied [21].

With recent advances in single molecule fluorescence imaging, it is possible to count mRNA and protein molecules in individual cells, especially in bacteria [13, 28]. In practice, only a few genes have been studied at a single-molecule single-cell level.

We report here a system-wide survey of mRNA and protein levels in single cells, achieved with single molecule fluorescent in situ hybridization (FISH), measurements of mRNA and single molecule imaging of yellow fluorescent protein fused to genes expressed at their native chromosomal position. We observe that noise initially scales inversely in proportion to protein abundance (intrinsic noise) and that extrinsic noise dominates for highly expressed proteins. Furthermore, we find that the copy numbers of mRNA and protein are uncorrelated in single cells.

2.3 Construction of a YFP-fusion library

E. coli does not generally recombine exogenous DNA into its genome. By expressing the lamda genes exo, beta and gam exogenously in *E. coli,* researchers were able to significantly improve recombination efficiency [6, 27]. This technology allowed the construction of a 857-gene *E. coli* library in which each strain was tagged with a sequential peptide affinity (SPA) tag on the C-terminus [3]. Ribosomal proteins were specifically excluded from this library. The library was used to find interacting proteins, and protein abundance was not measured. This was the only library where proteins in their native chromosomal positions were tagged.

Instead of making a fluorescent protein fusion library from scratch, we decided to convert the existing SPA tag library to a yellow fluorescent tag library. This is accomplished with one set of primers targeting the SPA sequence, and leaves a small scar sequence of ~40 bp (Figure 2.1). Yellow fluorescent protein (YFP) was chosen because

its fluorescence excitation is away from the cellular autofluorescence peak, improving sensitivity [26, 28]. The fast-maturing variant of YFP, venus, was selected for its short maturation time [20] to minimize any lag in observing newly synthesized proteins for live cell imaging experiments.

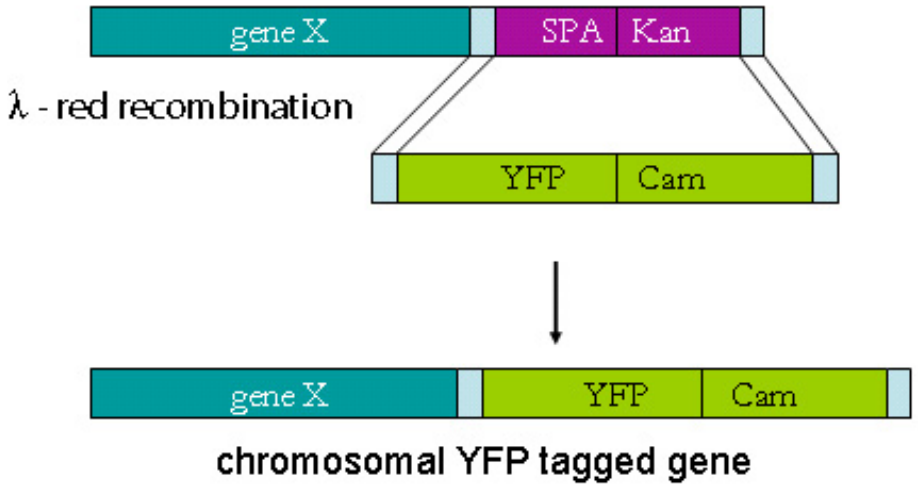


Figure 2.1 Construction of YFP library. This figure is adapted from Figure S10 from Taniguchi et al. [24] Starting with a library of *E. coli* strains that has genes tagged with a sequential affinity purification (SPA) tag on the C-terminus, the SPA-kanamycin (kan) region for each strain was replaced with a YFP – chloramphenicol (cam) cassette by λ-red recombination. Bacterial colonies that were cam-resistant were selected and sequenced to confirm the insertion.

Of the ~1400 strains in the SPA library that we converted (more genes were tagged after the publication of the paper in 2005), 1018 strains were confirmed by sequencing and showed no significant growth defects. Addition of the YFP tag is expected to be most disruptive to highly expressed proteins, lowering expression by ~30% (Supplementary information 2.12).

2.4 Imaging the cells and quantifying protein levels

To facilitate high throughput imaging of the bacterial cells for imaging, we designed a microfluidic chip that is compatible with a 3D translational stage. The chip

can hold 96 strains of bacteria immobilized to the poly-lysine coated coverslip in independent channels. ~4000 cells were imaged in fluorescence and phase contrast (to detect the outline of the cell) for each strain (Figure 2.2).

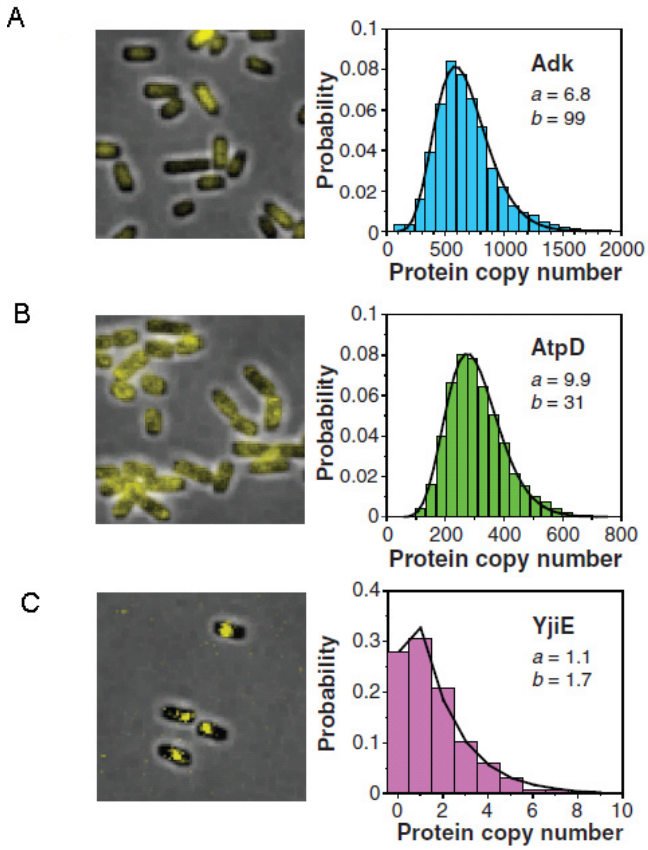


Figure 2.2 Sample images and histograms. This figure is adapted from Figure 1 in Taniguchi et al. [24]. Phase contrast images overlaid with fluorescence images (false color yellow) are shown. Protein localization can be inferred from the fluorescence images. The histogram of protein copy number per average cell volume (concentration) is shown for three proteins expressed at different levels. The histograms are fit with a gamma distribution, and the *a* and *b* parameters are given (explained in section 2.6). (A) Adk is a cytoplamic protein, evident from the diffused intracellular localization, that is expressed at very high levels. (B) AtpD is a membrane-bound protein, evident from how the fluorescence signal is not intracellular, that is expressed at high levels. (C) YjiE is a predicted DNA-binding protein, evident from the non-diffused intracellular distribution, that is expressed at low levels.

In previous single molecule experiments, the detection of the YFP was aided by the immobilization of the protein to membrane or DNA, permitting direct counting of proteins expressed at low levels [5, 8, 28]. Highly expressed proteins are generally not

suitable for such studies. For this study, we were not limited to immobilized or low copy proteins (Figure 2.2).

To obtain protein abundance from the images, the contribution of cellular autofluorescence was removed by deconvolution, and the absolute protein concentration was determined by calibration. We confirmed the single molecule sensitivity and calibration of our system with membrane-bound YFPs expressed at various levels, and also using purified fluorescent proteins (Supplementary Information 2.12).

2.5 Average protein abundance and noise

We calculated the average copy number (μ) and noise (σ^2/μ^2) for each protein (Figure 2.3). Average protein expression spans five orders of magnitude, ranging from 10^{-1} to 10^4 molecules/cell, with a peak at about 10 copies/cell. We also confirmed the range of protein expression by Western blot (Supplementary information 2.12).

Average protein expression in yeast spans about four orders of magnitude, from 10^2 - 10^6, with a peak at 2000 [12]. Correcting for cell volume (an *E. coli* cell is about 50 times smaller than a haploid yeast cell), the concentrations of proteins in *E. coli* and yeast are similar.

We noticed that the distribution of the average copy number of essential proteins in our dataset is different from that of the general dataset. Very few essential proteins (10%) are present at below 10 copies/cell, while half of all proteins are present at less than 10 copies/cell.

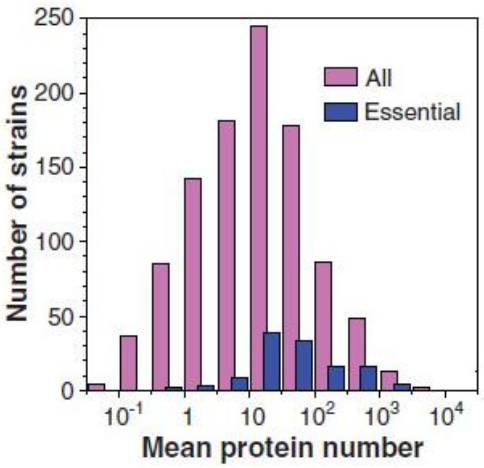
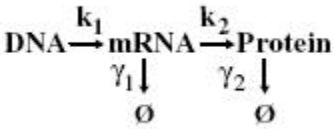


Figure 2.3 Histogram of mean protein number. This figure is adapted from Figure 2 in Taniguchi et al. [24] The histogram of essential proteins is plotted in blue. The distribution of essential protein copy numbers is different from the distribution of all proteins.

2.6 Understanding the population spread

To understand the distribution of protein in a population of cells, we consider the following kinetic scheme, which is the simplest model of gene expression



k_1 and k_2 are respectively the transcription and translation rate, while γ_1 and γ_2 are the mRNA and protein degradation rates respectively. Since proteins are mostly stable during exponential growth [15], γ_2 is dominated by the rate of dilution due to cell division for most proteins. The cell cycle in our experimental condition is about ~150 min. The lifetime of RNA is measured, and found to be on the order of a few minutes (2-5 min). Since the protein lifetime is much longer than RNA lifetime, fluctuations in the mRNA level can be integrated out. The number of protein bursts per cell cycle is given by $a = k_1/\gamma_2$, and the number of proteins per burst is $b = k_2/\gamma_1$. At low mRNA production where each protein burst is the result of one mRNA, a is the number of mRNA per cell cycle.

Assuming Poissonian mRNA production, and an exponentially distributed protein burst size, which were both experimentally observed [4, 28], the steady state distribution of protein copy numbers (x) is given by the gamma distribution

$$p(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}$$

which is defined by two parameters *a* and *b*. *a* and *b* are related to the mean (μ) and variance (σ^2) of the distribution such that $a = \mu^2/\sigma^2$ and $b = \sigma^2/\mu$. Of the 1018 protein copy number distributions, 1009 were well-fit by the gamma distribution. We will re-visit the interpretation of these fits later in this chapter.

2.7 Protein noise

Protein noise ($\eta^2=\sigma^2/\mu^2$, which is incidentally also 1/a) decreases with increasing protein abundance with a $1/\mu$ scaling until $\mu= 10$. The Fano factor ranges from 1 to 5. A Fano factor of 1 is consistent with a model where protein production and degradation are random events with a constant rate (Figure 2.4). For comparison, yeast shows a $1/\mu$ scaling with a Fano factor of 1200 for its mid abundance proteins [1].

We confirmed that the minimal model (described in Section 2.6) is sufficient to describe the production of low abundance proteins by tracking real-time protein production in single cells for several membrane-bound genes with low expression levels (Table 2.1). In general, the *a* and *b* parameters from the protein copy number distributions are similar to the directly observed burst frequency and burst size. The absolute differences are attributed to the different experimental conditions.

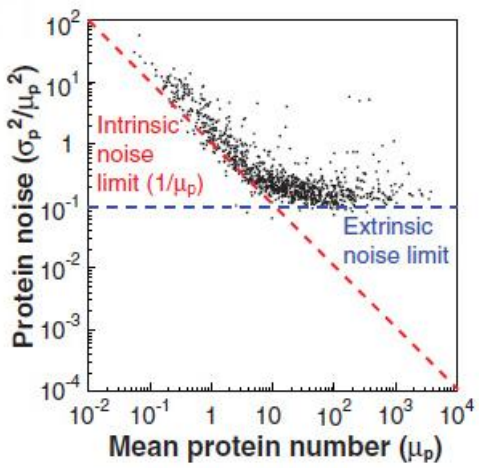


Figure 2.4 Relationship between mean protein number and noise. This figure is adapted from figure 2 of Taniguchi et al. Protein noise decreases in a 1/μ manner with increasing mean protein number (per cell) until μ=10, after which protein noise remains at 10%

strain	observed *a* (measured *a*)	observed *b* (measured *b*)
corA-YFP	11.4 (5.3 – 13.9)	3.7 (5.9 – 6.5)
ybdG-YFP	5.9 (2.5 – 5.5)	2.1(4.7)
ycjF-YFP	3.8 (1.8 – 5.6)	2.3 (5.2 – 5.6)

Table 2.1 Real-time burst frequency and burst size for 3 genes. This table is taken from table S4 of Taniguchi et al. [24]. For three membrane-bound proteins, we followed protein expression in real-time to directly measure protein burst frequency (*a*) and burst size (*b*). We also calculated *a* and *b* from the mean and variance of the protein copy number distribution.

As mean protein copy number increases beyond 10, there is no further decrease in the observed protein noise. This two-regime scaling has been observed in yeast with some differences. Noise in yeast plateaus at 1% while noise in *E. coli* plateaus at 10%. The noise floor implies that the burst frequency (a = 1/noise) per cell cycle is limited to 10. This is unlikely to be true, and most likely reflects that the simplest model of gene expression is insufficient to describe proteins expressed at high levels.

We hypothesized that the noise floor at 10% reflects extrinsic noise. We confirmed that there is a slowly fluctuating noise in protein levels that persists for longer than a cell cycle in for several highly expressed proteins (Figure 2.5). In addition, we

constructed thirteen pairs of randomly selected genes, labeling one with YFP (venus) and the other with RFP (mCherry) and confirmed that the expression levels of different genes in a single cell are correlated. The degree of correlation is consistent with the observed noise floor (Supplementary information 2.12).

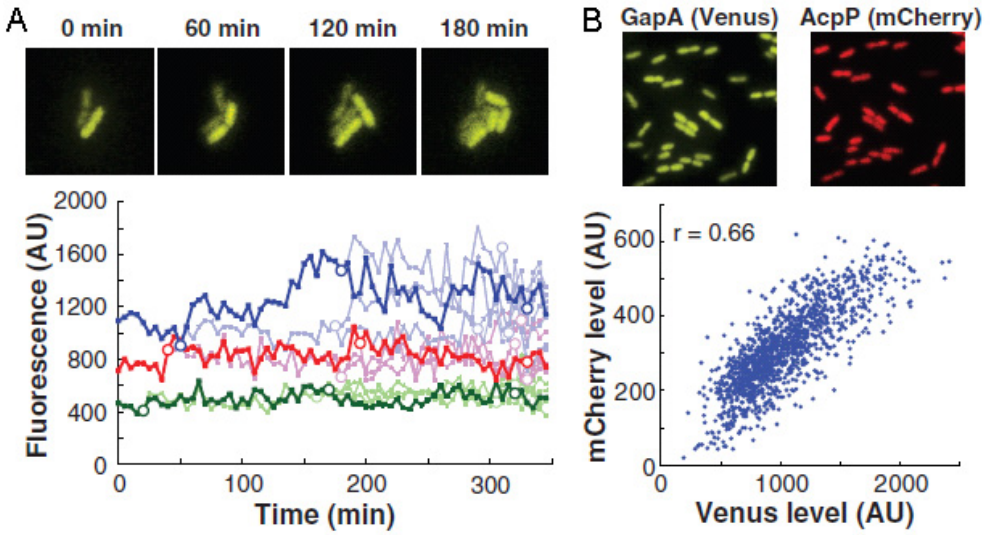


Figure 2.5 Confirming extrinsic noise. This figure is adapted from Figure 2 of Taniguchi et al. [24]. (A) Real-time observation of the slow fluctuation of protein levels. We follow various cell lineages of the strain AcpP-YFP, each lineage is colored by a different color in the trace. The dark line follows a single lineage, and the descendents are followed by lighter lines. Cell division events are indicated by a circle. The fluctuation within a cell over one cell cycle is small, confirming the existence of extrinsic noise. (B) Two-color measurements of two different proteins in the same cell. Two highly expressed proteins, GapA and AcpP, are respectively labeled with YFP and RFP in the same strain. The abundance of both proteins in a cell is plotted. The protein levels are correlated in single cells (r = 0.66), supporting the hypothesis that a global extrinsic noise dominates at high expression levels.

To account for extrinsic noise in our model for gene expression, we allow *a* and *b* to vary slowly with time such that the cell population no longer has a homogenous rates of mRNA and protein production. This still results in an apparent gamma distribution of protein copy numbers, explaining why most of the data was well fit even though extrinsic noise was not accounted for in the original model.

2.8 RNA counting: abundance and noise

We examined the abundance and noise of mRNA in single cells by fluorescence in situ hybridization (FISH) with single molecule counting. A single 20-mer oligonucleotide probe was designed against the yfp region of the mRNA, which allows us to detect all strains in the library without bias. We confirmed the accuracy of single molecule FISH by quantitative PCR (Supplementary information 2.12). The false positive rate of single molecule FISH is 0.1. The YFP and the Atto594 signal from the mRNA in the same cell can be simultaneously detected and spectrally resolved.

The mRNA abundance for 137 strains with high protein expression (>100 proteins/cell) were measured in single cells along with their protein abundance. The correlation between protein and mRNA is 0.77, consistent with previous observations [12]. The ratio of protein to mRNA ranges from 10^2 to 10^4. This is different from low copy proteins (<100 copies/cell) which have a protein burst size (protein/mRNA) ranging from 1 to 10. In comparison, yeast has a protein/mRNA ratio of about 1200 [1].

mRNA noise scales with the inverse of abundance (Figure 2.6). If mRNA production and degradation were stochastic with constant rates, we expect a Fano factor (σ^2/μ) of 1. The observed Fano factors range from 1 to 3, with a mode of 1.6, indicating that mRNA production and degradation are non-Poisson.

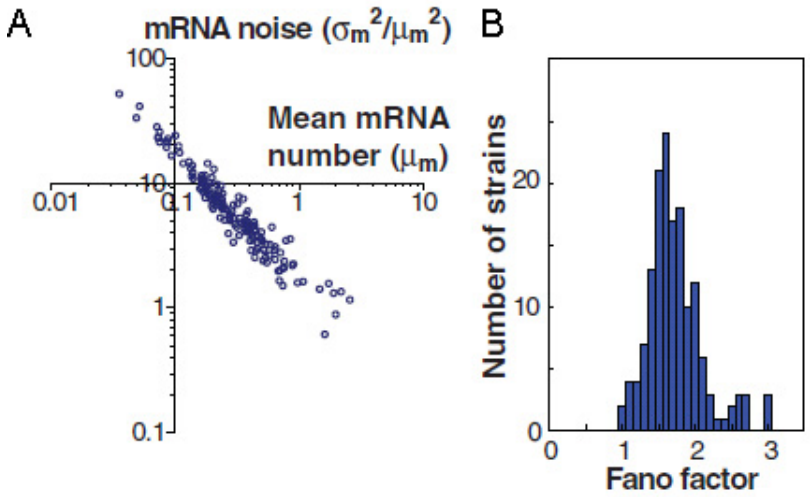


Figure 2.6 mRNA noise measured from single cells. This figure is adapted from Figure 3 of Taniguchi et al. [24]. (A) mRNA abundance (x axis) and noise (y axis) were measured from single cells by single molecule FISH. Noise scales proportional to $1/\mu$. (B) The Fano factor (σ^2/μ) for mRNA ranges from 1 to 3, and is centered at 1.6. In the case of Poissonian mRNA production and degradation, the Fano factor is 1.

2.9 Protein and mRNA abundance in single cells

We examined the correlation between protein and mRNA levels in single cells (Figure 2.7). For most of the highly expressed protein strains we surveyed, the correlation coefficients are close to zero, indicating that protein and mRNA in a single cell are generally not correlated.

Given that the average mRNA lifetime is less than 10 min (Figure 2.8), the mRNA copy number at an instant reflects the recent past. Because proteins are stable, the observed protein level is the sum of expression throughout the 150 min cell cycle. It is thus not surprising that mRNA and protein levels are poorly correlated at any one moment in a cell. In addition, a global noise in translation could also contribute to the lack of correlation between mRNA and protein levels at an instant.

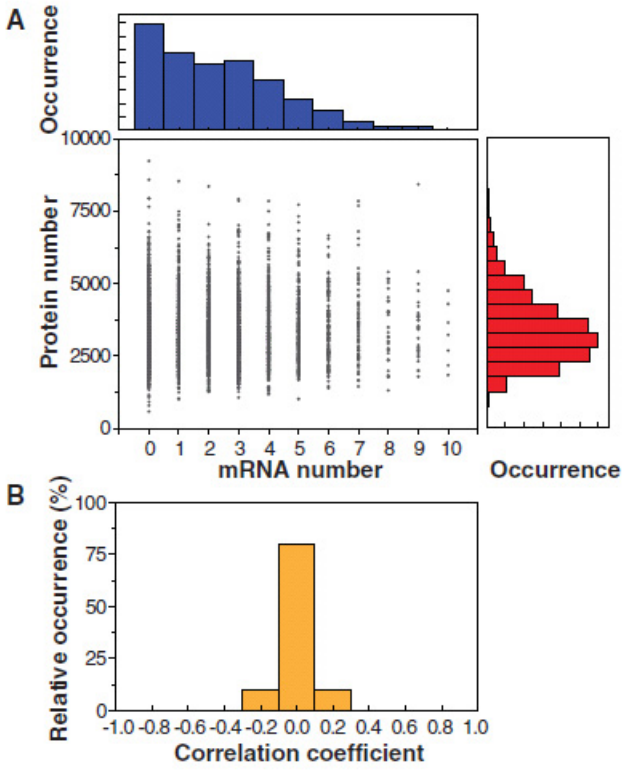


Figure 2.7 No correlation between mRNA and protein levels in a single cell. This figure is adapted from Figure 4 of Taniguchi et al. [24]. (A) The mRNA copy number, measured by single molecule FISH, and protein copy number for a single cell is plotted for the tufA-YFP strain. The correlation coefficient is 0.01. (B) Correlation coefficients for mRNA and protein in single cells for 129 strains with errors <0.1. The lack of correlation between mRNA and protein in single cells is general.

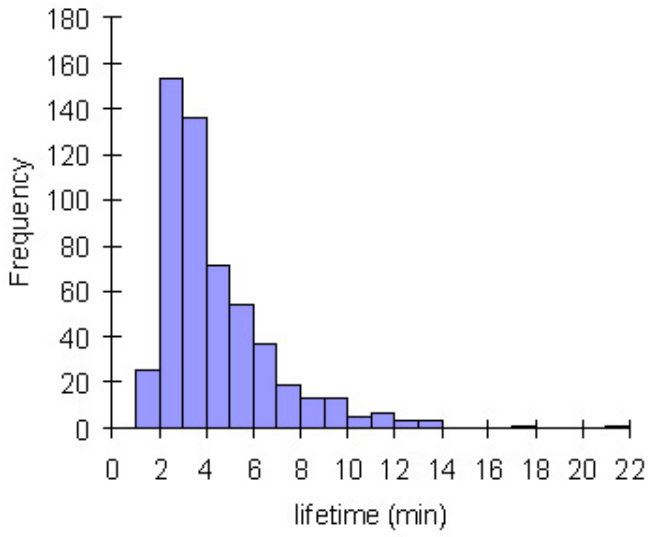


Figure 2.8 Histogram of RNA lifetimes. The lifetime of 541 genes was measured using a modified protocol from Bernstein et al. [2]. Most lifetimes are between 2 and 4 min, and the average is 4.4 min.

2.10 Conclusion

A yellow fluorescent-fusion protein library of 1018 genes expressed from their native loci was constructed for *Escherichia coli*. The fast-maturing YFP variant, venus, was chosen [20] so that the strains are suitable for real-time imaging studies on top of abundance and localization measurements.

The library strains were imaged using a microscope capable of single molecule detection. Besides measuring protein abundance in single cells, single molecule FISH was used to measure mRNA copy numbers. We were able to survey abundance and noise in proteins and mRNA in *E. coli* for a large number of natively expressed genes for the first time.

Protein expression ranges from 0.1 to 10,000 per cell, not including ribosomal proteins. At low expression, protein noise scales as $1/\mu$ reflecting intrinsic noise. Most

essential proteins are expressed at 10 copies/cell or more, which allows the cell to achieve minimal noise. At high protein abundance, we find that extrinsic noise prevails, and at least part of the extrinsic noise is from the translation process. This extrinsic noise is 10%, which is much larger than the 1% observed in yeast [21].

The noise structures of mRNA and protein are different. The mRNA of highly expressed proteins have a noise scaling proportional to $1/\mu$. Although mRNA and protein abundances are correlated for a population of cells, the correlation is almost zero in single cells. The large difference in the stability of mRNA and protein explains the poor correlation in single cells, but other factors like extrinsic translational noise are needed to explain the zero correlation.

Taken together, this study has provided a global view of gene expression, and its regulation in a population of *E. coli* cells, as well as in single cells.

2.11 Materials and methods

2.11.1 Construction of YFP-fusion library

Strains from an existing chromosomal fusion library (Butland 2005) were grown up to $O.D_{600}$ ~0.6 in LB from an overnight culture. The cells were grown at 30C at all relevant steps. The cultures were transferred to Mini-Tubes (Bel-Art F37857-0000) for a 15-minute, 42°C heatshock, and then chilled on ice. The cells from each strain were centrifuged at 3,800 rpm × 10 minutes at 4°C (Sorvall Super T21, ST-H750 rotor), to remove the supernatant and then washed twice with cold water (Millipore) before they were concentrated and transferred to 1 mm gap electroporation cuvettes (VWR International). The venus-chloramphenicol (CAM) resistence cassette was prepared with

Platinum PCR Supermix (Invitrogen) using primers VenF-SpaF (aactactgctagcgagaatttgtattttcagggtgagctcagcaagggcgaggagctgttcac) and CamR-KanR (ggcgtcgcttggtcggtcatttcgaaccccagagtcccgcctgccactcatcgcagtactgttgt) (Integrated DNA Technologies) and the PCR product was cleaned up using QIAquick PCR Purification Kit (Qiagen). Electroporation was performed at 1600V (BTX ECM399), and LB was added immediately.

Cells were allowed to recover for at least 3 hours before plating with chloramphenicol for selection. Colonies were screened for insertion by PCR. The PCR product from positive colonies was sequenced to confirm correct insertion using the Biopolymer Facility at Harvard Medical School.

2.11.2 Manufacture of the microfluidic chip

Photolithography and softlithography techniques were used to produce the microfluidic platform [19]. Poly-dimethylsiloxane (PDMS), a low-cost, optically-transparent silicon elastomer, was selected as the matrix of microfluidic chip. The chips were made by molding PDMS on a microfabricated silicon wafer. To prepare the mold, we designed a microfluidic pattern on AutoCAD 2004 (Autodesk Inc.) and output it into a photomask film using a commercial photoploting service (CAD/Art Services, Inc.) with a resolution of 20,000 dpi. We spin-coated an UV-curable epoxy (SU8-2025, Micro-Chem) with a 25 μm thick on a test-grade silicon wafer (University wafer). The designed microfluidic pattern was developed by exposing UV-light to the wafer through the photomask and immersing it in a developer solution. The PDMS was molded on the fabricated wafer by curing at 60°C in 45-60 minutes. φ0.75 mm holes were punched through the replicated PDMS sheet at the inlet/outlet positions. A coverslip (0.17 mm

thick, 48 × 60 mm, Brain research laboratories) and the PDMS sheet were treated by an oxygen plasma cleaner and were bonded each other. A microfluidic chip service provided some of the chips used in the experiments (Stanford Microfluidic Foundry).

2.11.3 Microscope

Single molecule fluorescent experiments were done on an inverted microscope (IX71, Olympus Americas, Inc.). Epi-illumination was provided by an Ar laser at 514 nm (Innova 300, Coherent) for Venus excitation and a fiber laser at 580 nm (VFL-P-Series, MPB Communications Inc.) for mCherry and Atto 594 excitation. Phase contrast illumination by a halogen lamp was also provided to identify the cell position and shape. Images were taken on an EM-CCD camera (Cascade 512B, Photometrics) with a 100 msec exposure through a 100× phase-contrast objective lens (NA = 1.35, Olympus). Samples were placed on a motorized 3D translational stage (MS2000, Applied Scientific Instrumentation). For real-time experiments, a temperature controller was attached to the sample (FCS2, Bioptechs). The light source was switched by mechanical shutters (VMM-D3, Uniblitz) and a dichroic mirror wheel. Automatic measurements were done by Metamorph software (Molecular Devices), which synchronizes the stage scanning, shutter control and camera acquisition.

2.11.4 Sample growth conditions

Cells were grown in LB media with 20 μg/ml Chloramphenicol and subsequently were inoculated into M9 media supplemented with 0.4 % glucose, amino acids and vitamin with 1:400 dilution. The cells were incubated at 30°C for 11-12 hours and were grown to OD_{600} = 0.1-0.5. To check that 4-5 cell divisions in M9 is sufficient to generate

cells in a steady state, we grew 20 randomly selected strains for >24 hours in M9 to confirm that the measured abundances were the same. Deep 2 ml 96-well plates (VWR) were used to culture many samples at once. During culturing, the plates were tightly capped and were placed on the side in a shaker to provide sufficient aeration. The doubling time was 150 minutes. Before imaging the cells were spun down in a tabletop centrifuge (Sorvall super T21, Kendro Laboratory Products) for 10 minutes at 3,800 rpm and washed once with 0.85% NaCl solution.

2.11.5 Sample preparation and measurement

The microfluidic chip we designed integrates 96 parallel independent channels and can hold 96 cell samples on a single coverslip. The measurements were automatically performed by scanning the microfluidic chip under a microscope capable of single molecule detection with a PC-controlled 3D translational stage. Samples with fewer than 500 cells, or that include many long unhealthy cells or lysed cells, were discarded and re-measured later. All samples were measured at least twice on two different days. The microfluidic chips used only once for the experiment.

A multi-channel pipette (12 channels, Rainin) was used to inject solution into the channels. For this purpose, the spacing between every 4 channel inlets was designed to match the spacing between pipette tips. The elasticity of PDMS works sufficiently to seal between the φ0.75 mm inlet and disposable plastic tips to inject the solutions. To immobilize bacterial cells on the microchannels, the channels were pre-coated with 0.1 % poly-L-lysine. After pre-coating, cells were injected into the channels and were incubated for more than 45 minutes for stable binding to the channel surface. Floating cells in the

channels were washed out by injecting 0.85 % NaCl solution, resulting in a single cell layer on the coverslip surface.

A combination of a phase contrast image and fluorescent image was taken at different positions along the channel profile. Typically, 10 sets of image were acquired per channel, resulting in 500-20,000 cells observation. To prevent over-saturated images, the camera gain is decreased for subsequent images if a very bright pixel was observed in the first image. The relationship between camera gain and obtained pixel value was been calibrated in advance. The elapsed time was 25 seconds per one channel. The recorded images were saved in 16-bit TIFF format.

As controls for microscopy measurement, fluorescein solution and plain NaCl solution in separate channels were imaged respectively. The former was used to compensate for the heterogeneity of laser illumination in the image field. The latter was used to subtract the dark noise count due to the EMCCD and the autofluorescence background from the microchannel and immersion oil.

2.11.6 Image processing

Automated image analysis was done by LabVIEW software (National instruments). Obtained fluorescence images were subtracted with the background image and were flattened using the fluorescein solution image. Phase contrast images were processed through a closed filter and a sharpen convolution filter and were thresholded to create binary images (Figure 2.9). The binary images were segregated into particles, which were filtered by an area, minor and major caliper lengths, in order to exclude overlapped cells, long unhealthy cells, cell debris and the other unexpected objects from the analysis. The integrated fluorescence intensity within the entire cell area was obtained

for each cell and was normalized by cell volume: (2/3)*(cell area)*(cell minor caliper length).

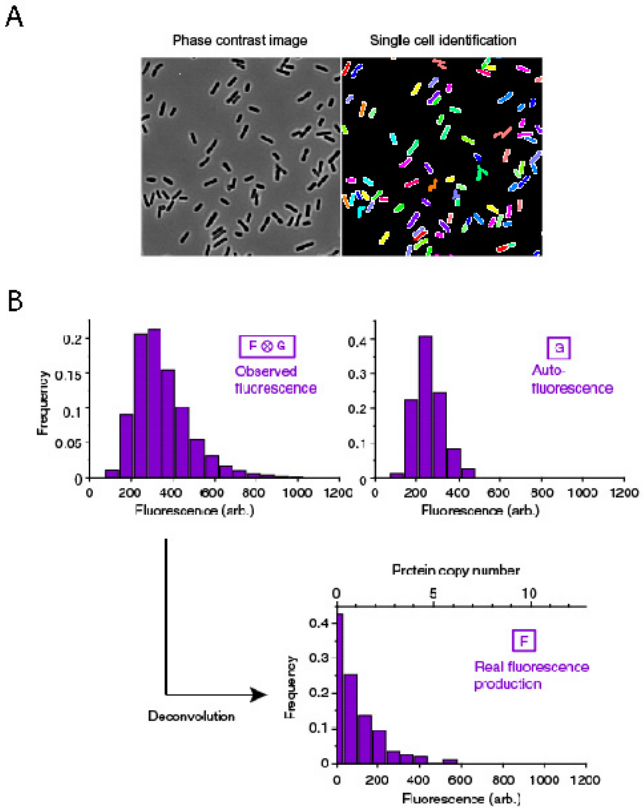


Figure 2.9 Image processing.This figure is adapted from Figure S14 from Taniguchi et al. [24] (A) Phase contrast images were reduced to binary images through image filters to find distinct particles that correspond to cells (the different colored objects in the right image). The particles were then filtered by particle size, area and shape to identify single cell boundaries (highlighted in white). The fluorescence count was integrated for the entire area for each cell. (B) The net protein copy number distributions (F) were calculated by deconvoluting the measured fluorescence histogram (F⊗G) with cell auto-fluorescence histogram (G).

2.11.7 RNA lifetime measurement

We modified the protocol used by the Cohen laboratory [2] to use RNAseq as a readout. Overnight cultures of the parent strain DY330 [27], was diluted 1:1000 into M9 with 0.4% glucose, vitamins, amino acids, 0.15ug/ml biotin, 1.5uM thiamine and grown to OD600nm = 0.3 at 30C. Rifampicin was added at a final concentration of 500ug/ml

and aliquots were removed at 0, 2, 4, 6, 8 minute time points and quenched in a 10% volume of cold phenol:ethanol (9:1). The cells were then harvested by centrifugation and washed once with 0.85% NaCl solution before storing in -80°C.

Frozen cell pellets were resuspended in 1 mg/ml lysozyme TE buffer and lysed by an equal volume of Cell Lysis Buffer (Purgene). The suspension was then extracted 1-2 times in 1 volume of acidic phenol/chloroform (OmniPur). The aqueous layer was extracted once in chloroform, and RNA was collected in RNA Clean and Concentrator columns (Zymo Research). Contaminating DNA was removed by DNase I (NEB) treatment for 30 min at 37°C, and the resulting RNA was repurified.

Starting with 5 μg of RNA, rRNA was removed first using Ambion's MICROBExpress following manufacturer's protocol, except RNA was collected using Zymo's RNA columns. A second rRNA removal step was performed following the protocol described in Affymetrix Expression Handbook, substituting enzymes MMLV (Ambion), RNase H (NEB), and DnaseI (Amplification grade, Invitrogen). 150-300 ng of RNA remain.

The purified RNA was fragmented using Ambion's Fragmentation Reagent at 70°C for 5 min, and collected by Zymo's RNA columns. RNA seq libraries were prepared according to Illumna's RNAseq protocol, using NEB enzymes and barcoded adapters (Integrated DNA Technologies). The DY330 libraries were pooled and sequenced with an Illumina GA II machine (Center for Systems Biology, Harvard University).

32 base reads were mapped to the W3110 genome using the Illumina Pipeline software. Further data analysis was performed using either homebrewed python2.6

programs, or simple Excel spreadsheets. The reads were then normalized by the stable gene ssrA [23]. Reads were sorted by genes, and linear regression of the log2(abundance) versus time point was used to calculate RNA half life. Fits with $R^2 > 0.7$ were reported.

In the following chapters, we discuss further considerations and corrections that should be made to the calculation of RNA lifetime. The conclusions made in this chapter based on the short lifetime of RNA relative to protein are not affected.

2.11.8 Real time observation of protein expression

Real-time observation of library strains were done as described previously [28]. Cells were centrifuged and placed between a 3% agarose gel pad and a glass coverslip. For low copy strains, the gel pad was made with M9 media supplemented with 0.4% glucose, 0.05 % casamino acid, 0.15 μg/ml biotin and 1.5 μM thiamine. For high copy strains, the gel pad was made with M9 media supplemented with 0.4% glucose, amino acids, vitamin and 0.2% casamino acid. The gel was set in an imaging chamber (FCS2, Bioptechs) and was kept at 30°C during observation. Image acquisition was done every 5 minutes for 5-9 hours. For low copy strains, a higher laser power (~600 W/cm2) was used to image single molecules with 50 ms exposure, followed immediately by 2 additional images to completely photobleach existing fluorophores. In contrast, for high copy strains, a lower laser power (~1.3 W/cm2) was used for the fluorescence excitation to prevent photobleaching. To exclude the frame-to-frame variation of fluorescence intensity due to the inconstancy of auto-focusing, shutter timing and a stage drift, we normalized the fluorescence values of cells by their average for each frame.

2.11.9 Fluorescence in situ hybridization (FISH)

The FISH probe is comprised of a 20-mer oligodeoxynucleotide (Venus495r, 5'-TCCTCGATGTTGTGGCGGAT -3') with a covalently linked a dye molecule (Atto 594) on the 5' end. The oligonucleotide sequence was chosen such that it is the reverse complement of a region on the *yfp* mRNA that has the least frequency of secondary structures. Atto 594 (Atto-tec GmbH) was chosen for its brightness, photostability, and the reduced nonspecific binding during in situ hybridization (data not shown). The dye is linked to the oligonucleotide via NHS ester reaction, followed by RNase-free HPLC purification (custom made by Sigma-Aldrich).

Library strains were inoculated and grown under the same condition stated above (Section 2.11.1). At ~ 0.3 OD, 950 µl of each cultured strain in the deep 2 ml 96-well plate was rapidly mixed with 950 µl of pre-chilled 2X fixation solution (7.4% formaldehyde and 2X RNase-free PBS in DEPC-treated water (Ambion)). The mixture was shaked vigorously briefly and incubated on ice for 15 min. The cells were then pelleted with a tabletop centrifuge (Sorvall Super T21) for 10 min at 3,800 rpm, and washed twice with ice-cold RNase-free PBS solution (Ambion). After the wash, the cells were resuspended in 70% ethanol and incubated at RT for 1 hour. Finally, the cells were spun down and washed with the Wash Buffer (25% formamide (Ambion) and 2X SSC (Ambion) in RNase-free water (Ambion)). The cells were resuspended in ~20 µl of the Wash Buffer.

The hybridization protocol was originally adapted from Femino et al [10], Maamar et al [18], and Zong et al [29]. The condition was further optimized for the YFP probe (Venus495r) in *E. coli*. We used only one single oligonucleotide probe with only

one dye molecule. This strategy offers an advantage of counting overlapping spots using either intensity or photobleaching steps. This is important for us when measuring the mRNA-protein correlation, because it requires the knowledge of the absolute copy number in each cell. It is still advantageous to use multiple probes in most cases, especially in bigger cells where single fluorophore cannot be easily detected.

The Hybridization Buffer consists of 25% formamide, 2X SSC, 10% dextran sulfate (Sigma Aldrich), 0.2 mg/ml BSA (New England Biolabs), 2 mM ribonucleoside vanadyl complex (Sigma Aldrich), and 0.1% *E. coli* tRNA (Sigma Aldrich). 10 µl of the cells prepared in the previous step was mixed with 50 µl Hybridization Buffer and 2.5 µl of 30 nM the FISH probes (Atto594-Venus495r) dissolved in RNase-free water and 0.2 mg/ml BSA. The hybridization mixture was incubated in a 96-well plate at 30°C for 9 hours.

Following the 30°C incubation, the cells were washed with the Wash Buffer twice, incubated in Wash Buffer at 30°C for 1 hour, and then washed once again with the Wash Buffer and with PBS once. The cells were resuspended in 10 µl PBS.

2.11.10 Imaging FISH samples

The hybridized and washed cells were immediately applied to poly-D-lysine (Sigma Aldrich) coated glass coverslips. The coverslips (25 mm × 75 mm, Belco) were cleaned and prepared as follows: 30 min sonication in 1 M potassium hydroxide, 30 min sonication in purified water (Millipore), 1 min blow dry by nitrogen gas, 10 min in plasma sterilizer, 40 min in 0.03 % poly-D-lysine (Sigma Aldrich), 1 min rinse in deionized water, and 1 min blow dry by nitrogen gas.

Each coverslip is then adhered to a 16-well silicone gasket (FlexWells, Grace Bio-labs). The cells were allowed to adhere to poly-D-lysine coated coverslip for 30 min while covered to prevent evaporation. Each well is then washed extensively with RNase-free PBS. After the final wash, each well is filled with 140 mM 2-mercaptoethanol (Sigma Aldrich) in PBS, and covered with a cleaned glass slide.

Single molecule imaging was performed with the same microscope used for protein imaging. For Atto 594 imaging, a 580 nm fiber laser (MPB Communications Inc.) was used. An achromatic quarter waveplate (Thorlabs) was used to create near-circular polarization at the objective imaging plane. The fluorescence filter set includes an excitation filter (HQ575/50X, Chroma), a dichroic mirror (z594rdc, Chroma), and an emission filter (D635/55M, Chroma). The laser intensity at the image plane is ~100 W/cm2. Each image was recorded in 1 s. For YFP imaging, a 514 nm laser (Innova 300, Coherent) was used. The filter set includes an excitation filter (D510/20X, Chroma), a 525 nm longpass dichroic mirror, and an emission filter (HQ545/30M, Chroma). The laser intensity at the image plane is ~100 W/cm^2. Each image was recorded in 100 ms. No statistically significant crosstalk was observed between the YFP channel and the Atto 594 channel.

Automated image acquisition using Metamorph (Molecular Devices) allows sequential imaging of each 16-well coverslip, as described earlier. YFP and FISH images were recorded for 20-30 field-of-views for each strain, with an average of ~1000 cells total. Images without laser excitation were recorded to serve as offsets of the actual fluorescence images. Images of dilute dye solutions were recorded to correct for the slight inhomogeneity of the field-of-view.

2.11.11 Image processing of FISH data

The phase-contrast images and the YFP images were analyzed as described in the earlier section. The FISH images were subtracted with camera offset, and the field-of-view was flattened using the image of dilute dye solution described in the previous section. Fluorescence spots corresponding to localized mRNA were identified with a peak-searching algorithm written in Matlab (The MathWorks). The algorithm searches for pixels that have both (i) pixel intensity above a pre-defined threshold and (ii) image curvature above a pre-defined threshold. The thresholds are adjusted so that all fluorescent spots are identified via visual inspection, and that all identified peaks correspond to actual spots. For each fluorescent spot, the fluorescence intensity above background was calculated in the 5-by-5 pixels (corresponding to 800 × 800 nm) surrounding the peak. If more than one peak are identified within the 5-by-5 region in a same cell, the masks are merged so that each pixel is counted only once. For each cell, the following information were recorded: the total FISH signal, the total YFP signal, the size of the cell, and the lengths of the major and minor axes of the cell. The accuracy of the analysis method is illustrated in the following sections, where we discuss the false-negative and false-positive rates of the assay.

To reduce the gene dosage effect on the mRNA copy number distribution, a small set of cells within a certain size range was used for further analyses. The area of the cells ranges from ~1.9 μm^2 to ~4 μm^2, depending on the stage of the cell cycle. We selected the cells whose sizes are between 1.92 μm^2 to 2.30 μm^2. The fluorescence signal histograms are computed for each strain, including a mock strain which contains no YFP

gene. The resulting histograms are deconvolved from the histogram of the mock strain, which represents the nonspecific signal level. The fluorescence signal is then normalized to the signal from a single fluorophore to convert to the absolute number of probes. The 95% confidence level in determining the mean fluorescence level, the Fano factor of the distribution, and the mRNA-protein correlation were estimated by bootstrapping.

2.11.12 Ensemble mRNA copy number measurement

To independently confirm the fidelity of FISH measurement, we compared the average mRNA copy number per cell measured by FISH with the number measured by quantitative PCR in bulk. The comparison was done in the *E. coli* strain PC2a, in which the YFP expression is control under the *lac* promoter at the *lac* operon locus on chromosome. The bulk measurement is performed in courtesy of Professor Nam-Ki Lee, with the following procedures.

The cells were grown overnight in M9 glycerol medium supplemented with amino acids and vitamins at 37C. On the next day the culture was diluted 200 times into fresh M9 glycerol medium supplemented with amino acids, vitamins, and 1 mM IPTG. When the O.D. of the culture reached 0.2-0.3, 300 μl of the culture was transferred into 600 μl bacterial RNA stabilizer solution (Invitrogen). Total RNA was extracted using the RNeasy kit (invitrogen) following the manufacturer's guide. The residual gDNAs was removed using the TurboDNA-free kit (Ambion). The cell density of the culture was measured using a cell counter (Hausser Scientific). Reverse transcription was performed using SuperScript III (Invitrogen) at 50°C for 1 hour with primer sequence 5'-CGTCGTCCTTGAAGAAGATGG. Quantitative PCR was performed using the 7500

Fast Real-Time PCR System (Applied Biosystems), with a Taqman probe (5'-(FAM)ATCGCCCTCGCCCTC(MGB)) and two primers (5'-CGTCGTCCTTGAAGAAGATGG and 5'–CCGACCACTACCAGCAGAACA). Calibration was done using mRNA generated by *in vitro* transcription. The *E. coli* strain BW25993 was harvested, and mixed with known amount of *venus* mRNA at the amount of 1.0×10^{10}, 1.0×10^{9}, 1.0×10^{8}, and 1.0×10^{7} molecules, respectively. We extracted mRNA three times independently and for each extracted mRNA six RT-PCR reactions were performed. The calibration was performed in parallel with each measurement.

2.12.13 Western blot

Cells were grown up at 30°C in M9 glucose with amino acids, vitamins and biotin. After pelleting, the cells were lysed in 4% sodium dodecyl sulfate. Protein loading buffer was added and the sample was heated at 100°C for 5 min. Samples were run on a Ready Gel 4-15% Tris-HCl gel (BioRad), and transferred to nitrocellulose membrane (BioRad) using wet transfer. Blocking was performed in a 0.1% Tween 20 PBS solution with 5% milk and was followed by overnight incubation with an anti-GFP-tag antibody (1:5000, A6455 Invitrogen) at 4°C. HRP conjugated anti-rabbit antibody (sc-2955, Santa Cruz Biotechnology) was used at 1:2500 for secondary incubation. The blot was visualized using chemiluminescent substrates (SuperSignal West Femto Substrate, Thermo Scientific), and captured on film (Biomax Light Film, Kodak).

2.12 Supplementary information

2.12.1 Calibration of single molecule fluorescence

The fluorescence count corresponding to single molecule fluorescence was calibrated in two ways; one is a single molecule method (i) and the other is a bulk method (ii). We confirmed that the values from these two methods are consistent with each other.

(i) We measured SX4 strain expressing membrane-bound Tsr-Venus [28]. We obtained fluorescence counts from single Venus molecules by measuring the localized fluorescent spots. The obtained count was 161 counts/molecule/average cell volume. The cell volume is an average over the population.

(ii) We purified Venus protein and compared its fluorescence counts with that of a culture of a library cell strain (AcpP-Venus). The concentration of purified Venus ([*Venus*]) was measured by fluorometer (DU800, Beckman Coulter), and the density of cells (*C*) was obtained by a cell counter (Hasser Scientific Partnership). We injected cells, purified Venus, and 0.85% NaCl solution into different microfluidic channels pre-coated with BSA, and their fluorescence was observed with a 10× objective lens (Olympus). Comparison of fluorescence counts between those channels gives the protein number per cell for the cell sample, *n*, by the equation:

$$n = \frac{[Venus]\dfrac{F_c - F_w}{F_v - F_w} N_A}{C}$$

where N_A is the Avogadro's number, and F_c, F_v and F_w are the fluorescence counts from cells, Venus protein and NaCl solution, respectively. By comparing the value with the steady-state library data, we obtained a calibration count consistent with the value determined in (i). The average fluorescence per cell for AcpP was 741,223, calculated as described in chapter 2.11; this gives 141 counts/molecule/average cell volume.

2.12.2 Consistency with single molecule counting

Instead of counting protein molecules by detection by localization [28], we used a deconvolution method to determine the protein count as described in the supplementary methods. The deconvolution method has several advantages: it can measure both low and high copy proteins, and is not affected by protein localization. To check if the deconvolution method is consistent with the results from localized single-molecule counting, we measured the abundance of localized Tsr-YFP molecules in strain SX701 [5] as a function of an inducer, TMG. We found that the induction kinetics of tsr-venus is the same whether measured by counting localized molecules or by deconvolution (Figure 2.10). This indicates that the deconvolution method has enough sensitivity and resolution to detect similar changes as the localization method.

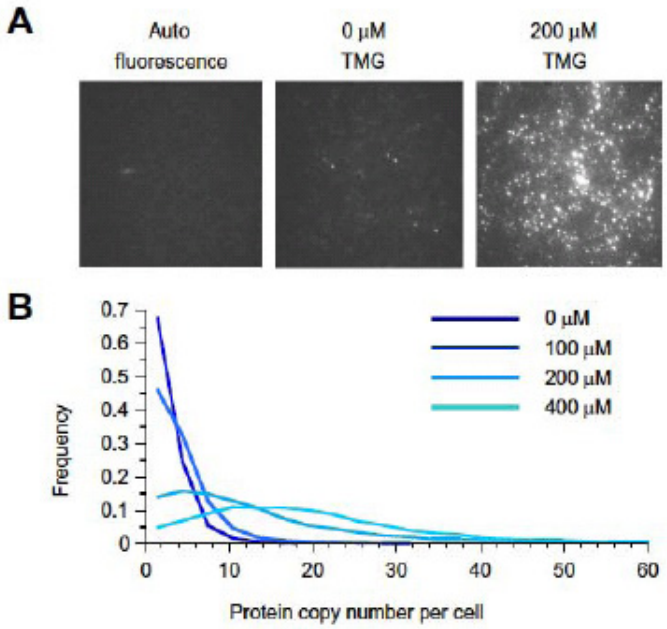


Figure 2.10 Protein abundance by deconvolution. This figure is adapted from Figure S2 of Taniguchi et al. [24]. (A) Images of SX701 cells taken at different levels of TMG induction. The control cells, BW25993, do not contain YFP (auto fluorescence). (B) Histogram of protein abundances determined by deconvolution agrees with previous counts of protein by single molecule localization [5].

2.12.3. Detection limit of the measurement system

We first confirmed the single molecule sensitivity of our microscope by observing a one-step photobleaching of low-copy, membrane-bound library strains. We checked the detection limit of our measurement system by deconvoluting the fluorescence histogram of cells with no YFP from the auto-fluorescence histogram of the control cells. The limit was determined to be 0.08 ± 0.08 /cell (mean ± SD, $N = 15$), which allows detection of 99.3% of the library data. We also verified the accuracy of our measurements by showing that measurements made on two separate occasions are largely reproducible ($r = 0.92$). We also determined the detection limit of protein noise to be ~0.01 by measuring control samples.

2.12.4 Perturbation of SPA-venus tag to native protein expression

We use a reporter protein, LacZ, to check that fusing SPA-venus (SPA is a scar sequence from the previous library) or venus to proteins did not affect their expression levels. Assuming that the β-galactosidase (β-gal) activity of LacZ is unchanged by C-terminal protein fusions, we can use the Miller assay to report on LacZ protein abundance. We can vary inducer (IPTG or TMG) concentration to achieve different levels of LacZ expression, to mimic low or high copy proteins. The Miller assay was performed using the yeast β-galactosidase assay kit (Pierce). We measured three different constructs: 1) LacZ, 2) LacZ + Venus, 3) LacZ + SPA + Venus. The C-terminal fusions of Venus and SPA-Venus resulted in at most a 2-3 fold reduction at high expression levels of LacZ as measured by β-gal activity (Figure 2.11). There was less perturbation in protein expression when LacZ protein was at low concentrations. This suggests that the C-terminal SPA-venus tag did not cause significant changes in protein expression levels.

We also confirmed that the β-gal activity is proportional to the observed fluorescence value (Figure 2.11C). This shows that there is no appreciable self-quenching of fluorescence at high expression levels. Furthermore, the direct comparison between the fluorescence and β-gal activity shows that the fluorescent reporter is linear for at least three orders of magnitude.

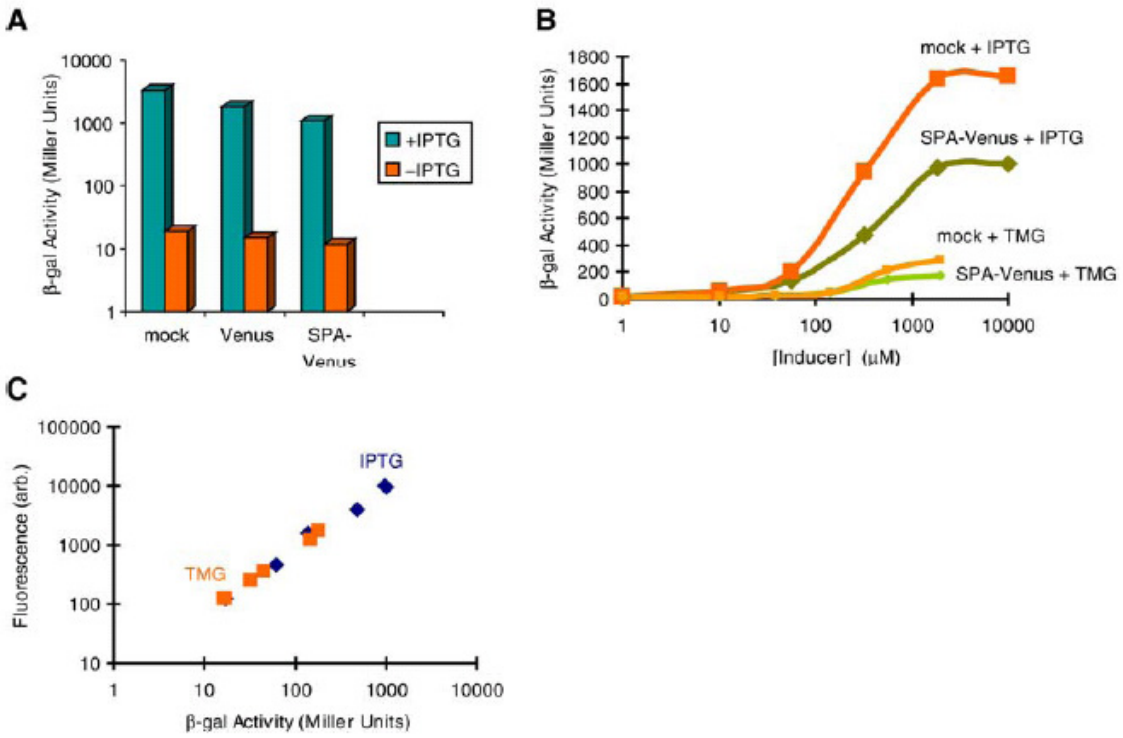


Figure 2.11 Checking perturbation by the YFP tag. This figure is adapted from Figure S4 of Taniguchi et al. [24]. (A)The change in lacZp abundance, reported by β-galactosidase activity, with and without 1mM IPTG for three strains containing lacZ, lacZ-venus, and lacZ-SPA-venus. (B) The dependence of lacZp abundance, reported byβ- galactosidase activity, as a function of concentrations of inducers, IPTG and TMG. The tag perturbs lacZ expression most at high expression levels. (C) There is a linear relationship between lacZ abundance measured by β-galactosidase activity and the fluorescence level in the LacZ-SPA-venus construct.

2.12.5 Consistency of fluorescence abundance measurement with other methods

In this work, we have shown that average protein abundances span five orders of magnitude. We note that this is consistent with previous observations. Previously a mass spec and western blot analysis [17] has shown an abundance data for high copy genes in an E. coli strain (K-12, MOPS media), and has indicated that the average abundance spans 10^3-10^4 magnitude, which is largely consistent with our result. We confirmed that their data sets, despite the different growth condition and strain background, are largely correlated with our data (r = 0.58 and 0.48, respectively), especially for high copy numbers. In addition, for a specific gene (LacZ), we have already confirmed that the YFP fluorescence is linearly proportional to its enzymatic activity at different expression levels on 2-3 orders of magnitude.

Further, we have performed a western blot analysis, using anti-GFP antibody, of several genes with different expression levels, including Adk and YjiE shown in Figure 2.2. We can observe at least 1000-fold differences in protein expression levels. Highly expressed genes (AcpP, 5,000 copies/cell, and CspC, 8,000 copies/cell) can be easily detected at dilutions of 100-1000 fold (dilution factors shown in brackets where applicable). Medium-high expressed genes (GyrA, 88 copies/cell, Adk, 600 copies/cell) can also be easily detected by the antibody. YjiE (5 copies/cell) is barely visible in the blot (band indicated by the asterisk), but genes expressed at less than a few copies per cell (DapA, 0.6 copies/cell, HypE, 0.2 copies/cell) are not distinguishable from the non-specific background (Figure 2.12).

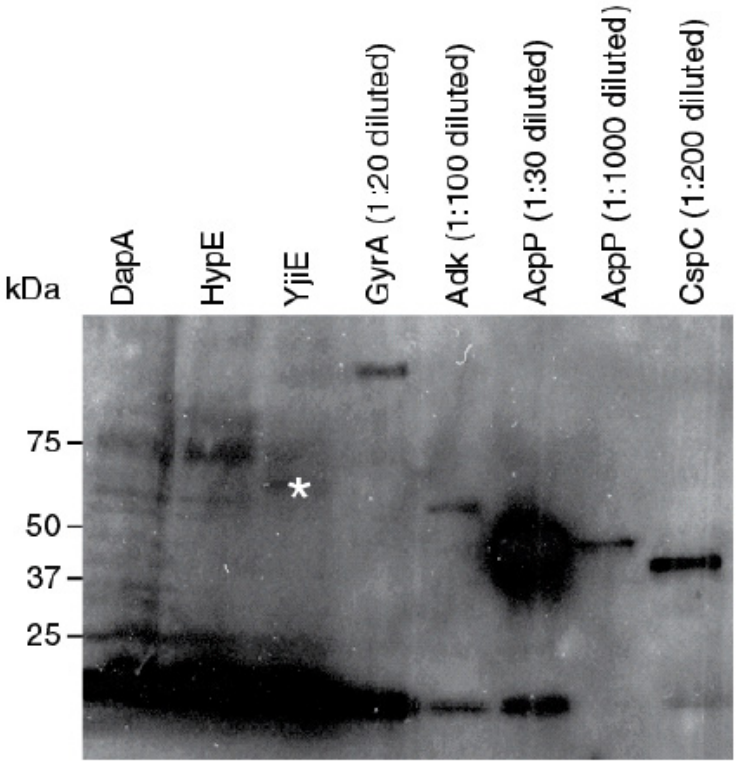


Figure 2.12 Western blot analysis of genes with different expression levels. This figure is taken from Figure S23 of Taniguchi et al. [24]. Genes are arranged in order of increasing expression levels. The expected sizes of the fusion proteins are 61 kDa (DapA), 60 kDa (HypE), 64 kDa (YjiE), 126 kDa (GyrA), 57 kDa (Adk), 38 kDa (AcpP), and 37 kDa (CspC). A white asterisk marks the band for YjiE. We used the nonspecific band at 10 kDa as loading control.

2.12.6 Correlation between two proteins in a single cell confirms global noise measurement

The correlation between two proteins in a single cell can give the value of the extrinsic noise (derivation in Taniguchi et al.[24])

$$\frac{\langle n_x n_y \rangle}{\langle n_x \rangle \langle n_y \rangle} = 1 + \eta_G^2$$

where $\eta_G{}^2$ is the global factor noise.

We observed positive correlations (r = 0.2-0.8) for all thirteen randomly selected two-protein combinations of highly expressed gene, where one gene is probed by Venus and the other is probed by mCherry. This confirms the existence of a global noise factor. We found that the normalized correlation factors, $\langle n_x n_y \rangle / \langle n_x \rangle \langle n_y \rangle$, was very uniform for all measured strains ($\langle n_x n_y \rangle / \langle n_x \rangle \langle n_y \rangle$ = 1.09 ± 0.03, mean ± SD, 13 strains), supporting the gene-independent property of the extrinsic noise. Thus, the global noise factor was determined to be 0.09. This means that the robustness of any gene function must take into account the unavoidable 30% variation in expression levels. The determined extrinsic noise limit is consistent with the limiting value of 10% protein noise obtained in our system-wide measurement.

2.12.7 False positive and false negative detection in single molecule FISH

To determine the false-positive rate (including nonspecific hybridization) in our assay, the same FISH method and analysis were applied to the E. coli strain BW25993 [6] which contains no YFP coding sequence. One hybridization event was detected in an average of ten cells, indicating a false-positive rate of 0.1 per cell. Similarly, same false-positive rate was detected in an *E. coli* strain that contains the YFP coding sequence and in which the expression of the *yfp* mRNA was suppressed under the *lac* promoter (strain PC2a). The mRNA expression level in this strain was determined to be < 0.04/cell [28], which is below our false-positive rate. The fact that we observed the same false-positive rate in the wild type strain and in the strain that contain the YFP coding sequence but no *yfp* mRNA indicates that there is minimal hybridization between the FISH probe and the genomic DNA under the experimental conditions. The intensity distribution of the false-

positive signals resembles that of a single fluorophore, suggesting that the signal arose from actual probes, rather than noise in the imaging system, that are nonspecifically bound in the cell.

To determine the false-negative rate for our assay, we compared the mRNA copy number per cell measured by FISH verses that measured by quantitative PCR in bulk. *E. coli* strain PC2a, in which YFP is induced under the *lac* promoter, was grown in M9 glycerol minimal medium supplemented with 1 mM IPTG, amino acids, and vitamins, and was harvested in log phase for the proceeding FISH, RNA extraction, or cell counting. The average *yfp* mRNA copy number per cell measured by quantitative PCR is 4.1 ± 0.7 (SEM, $N = 6$), whereas that measured by FISH is 3.77 ± 0.07 (SEM, $N = 6{,}528$). Therefore, the detection efficiency of our FISH assay is ~92%.

2.13 References

1. Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., Barkai, N. (2006) "Noise in Protein Expression Scales with Natural Protein Abundance." *Nat. Gen.* **38**, 636 – 642

2. Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S., Cohen, S.N. (2002) "Global Analysis of mRNA Decay and Abundance in *Escherichia coli* at Single-Gene Resolution Using Two-Color Fluorescent DNA Microarrays." *Proc. Nat. Ac. Sci.* **99**, 9697 – 9702

3. Butland, G., Peregrin-Alvearez, J.M., Li, J, Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., Emili, A. (2005) "Interaction Network Containing Conserved and Essential Protein Complexes in *Escherichia coli*." *Nature* **433**, 531 – 537

4. Cai, L., Friedman, N., Xie, X.S. (2006) "Stochastic Protein Production in Individual Cells at the Single Molecule Level" *Nature* **440,** 358 -362

5. Choi, P.J., Cai, L., Frieda, K., Xie, X.S. (2008) "A Stochastic Single Molecule Event Triggers Phenotype switching in a bacterial cell." *Science* **322**, 442 – 446

6. Datsenko, K.A., Wanner, B.L. (2000) "One-step Inactivation of Chromosomal Genes in *Escherichia coli* K-12 Using PCR Products." *Proc. Nat. Ac. Sci.* **97**, 6640 – 6645
7. Dubnau, D. and Losick, R. (2006) "Bistability in Bacteria." *Mol Microbiol.* **61**, 564 – 572

8. Elf, J., Li, G.-W., Xie, X.S. (2007) "Probing Trancription factor dynamics at the single molecule level in a single cell." *Science* **316**, 1191 – 1194

9. Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S. (2002) "Stochastic Gene Expression in a Single Cell." *Science* **297**, 1183 – 1186

10. Femino, A.M., Fay, F.S., Fogarty, K. Singer, R.H. (1998) "Visualization of single RNA Transcripts in situ" *Science* **280**, 585 – 590

11. Friedman, N., Cai, L., Xie, X.S. (2006) "Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression." *PRL* **97**, 168302

12. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S. (2003) "Global Analysis of Protein Expression in Yeast." *Nature* **425**, 737 – 740

13. Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C. (2005) "Real-Time Kinetics of Gene Activity in Living Cells." *Cell*, **123**, 1025 – 1036

14. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O'Shea, E.K. (2003) "Global Analysis of Protein Localization in Budding Yeast." *Nature* **425**, 686 – 691

15. Koch, A.L., Levy, H.R. (1955) "Protein Turnover in Growing Cultures of *Escherichia coli*" *J. Biol. Chem.* **217**, 1 – 7

16. Kussell, E., and Leibler, S. (2005) "Phenotypic Diversity, Population Growth and Information in Fluctuating Environments", *Science* **309**, 2075 – 2078

17. Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M. (2005) "Abosolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation." *Nat. Biotech.* **25**, 117 – 124

18. Maamar, H., Raj, A., Dubnau, D. (2007) "Noise in Gene Expression Determines ell Fate in Bacillus subtilis." *Science*, **317**, 526 – 529

19. McDonald, Whitesides, G. (2002) "Poly(dimethylsiloxane) as a Material for Fabricating Microfluidic Devices." *Acc. Chem. Res.* **35**, 491 – 499

20. Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., Miyawaki, A. (2002) "A Variant of Yellow Fluorescent Protein with Fast and Efficient Maturation for Cell-Biological Applications." *Nat. Biotech.* **20**, 87 – 90
21. Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., Weissman, J.S. (2006) "Single-cell Proteomic Analysis of *S. cerevisiae* Reveals the Architecture of Biological Noise." *Nature* **441**, 840 – 846

22. Raser, J.M., O'Shea, E.K. (2004) "Control of Stochasticity in Eukaryotic Gene Expression." *Science* **304**, 1811 – 1814

23. Ray, B. K., and D. Apirion. (1979) "Characterization of 10S RNA: a new stable RNA molecule from Escherichia coli." *Mol. Gen. Genet.* **174**, 25-32.

24. Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., Xie, X.S. (2010) "Quantifying E.coli proteome and Transcriptome With Single Molecule Sensitivity in Single Cells." *Science* **329**, 533 – 538

25. Thattai, M., van Oudenaarden, A. (2001) "Intrinsic Noise in Gene Regulatory Networks." *Proc. Nat. Ac. Sci.* **98**, 8614 – 8619

26. Xie, X.S., Choi, P.J., Li, G.-W., Lee, N.K., Lia, G. (2008) "Single Molecule Approach to Molecular Biology in Living Cells." *Annu. Rev. Biophys.* **37**, 417 – 444

27. Yu, D., Ellis, H.M., Lee, E.-C., Jenkins, N.A., Copeland, N.G., Court, D.L. (2000) "An Efficient Recombination System for Chromosome Engineering in *Escherichia coli*." *Proc. Nat. Ac. Sci.* **97**, 5978 – 5983

28. Yu, J, Xiao, J., Ren, X, Lao, K., Xie, X.S. (2006) "Probing Gene Expression in Live Cells, One Protein at a Time." *Science* **311**, 1600 – 1603

29. Zong, C., So, L.-H., Sepulveda, L.A., Skinner, S.O., Golding, I. (2010) "Lysogen Stability is Determined by the Frequency of Activity Bursts from the Fate-Determining Gene." *Mol. Sys. Biol.* **6**, 440

Chapter 3: The dynamics of RNA synthesis and degradation

Contributions:

The project discussed here was first conceived as an RNA degradation project following discussion with Dr. Paul Choi, who helped with the experiments and data analysis of the RNA degradation measurement in the previous chapter. With encouragement from Sunney, Dr. Katsuyuki Shiroguchi conducted further analysis of that dataset and observed the two phases in the data. Dr. Shiroguchi and I realized that we were seeing elongating RNA polymerases in the data, and broadened the project to involve RNA elongation. I performed the bulk of the experiments, with assistance from Dr. Shiroguchi for the streptolydigin experiments. Most of the primary data analysis was performed by Dr. Shiroguchi. I contributed to sorting and mapping reads obtained from the Illumina machines, and performed secondary data analyses to find correlations.

3.1 Abstract

The dynamics of RNA production and degradation are important for understanding gene regulation, but there have been few system-wide measurements of these processes. Using RNA-seq, we obtained quantitative measurements of RNA abundance and degradation rates at sub-genic resolution genome-wide in *Escherichia coli*. We developed an improved method for accurately measuring RNA degradation rates, and found that degradation rates show little position dependence, in contrast to previous reports. Our findings may help distinguish between different models of RNA degradation. In addition to the degradation rate, our experiment simultaneously provides a novel

method of measuring the elongation rate of any native promoter. Thus, we also report the first system-wide measurement of RNA elongation rates. From these measurements of RNA dynamics, we calculated RNA synthesis rates to estimate the allocation of RNA polymerases in the genome. By further combining the protein measurements from Chapter 2, we estimated the stoichiometry of the transcriptional and translational machinery for different genes. Lastly, we analyzed the coordination of different RNA dynamics and predict co-transcriptional degradation for some transcripts.

3.2 Introduction

In *E. coli*, mRNA typically has a lifetime of minutes [3] (Bernstein 2003). The short lifetime means that there must be constant and active synthesis of RNA to maintain steady-state abundance. The ability to describe the dynamics of RNA synthesis and degradation is useful for understanding gene expression. The parameters *a*, the number of protein bursts per cell cycle, and *b*, the proteins per burst, which were used to describe the distribution of proteins in single cells in chapter 2, are determined in part by RNA dynamics. *a* is related to the frequency of mRNA transcription, and the number of proteins per transcriptional burst, b, is determined by the lifetime of the mRNA.

RNA dynamics can be directly tracked in live *E. coli* cells by tagging the mRNA with MS2 protein-binding sites, and fusing green fluorescent protein (GFP) to the MS2 coat protein that binds the RNA [9]. However, this method protects the RNA from degradation, prolonging its lifetime [9]. Moreover, only one mRNA can be studied at a time.

RNA abundance and degradation rate can be measured genome-wide by microarray, using just one probe to each gene [3]. Because steady-state abundance is the

balance between synthesis and degradation, we can calculate RNA synthesis rates from the abundance and degradation rate. A more sophisticated microarray experiment, using five probes to each transcript, found that after stopping RNA synthesis, the 5' end of RNA is always less abundant than the 3' end, suggesting that degradation generally proceeds in a 5' to 3' direction [20].

Next generation sequencing method, RNA-seq (Chapter 1), is increasingly the tool of choice for measuring RNA abundance genome-wide, replacing microarrays. Because RNA-seq detects by sequencing instead of hybridizing a probe, the experiment is not limited to a pre-defined set of probes. Background hybridization is also no longer a problem, thus the number of reads mapping to a gene is directly proportional to the abundance of the gene in the sample. As a result, RNA-seq data should be both richer and more quantitative. We thus took advantage of the improvement in technology to study RNA synthesis and degradation in more detail.

3.3 Measurements of RNA degradation using rifampicin need to account for residual polymerase activity.

Following the addition of rifampicin (time = 0 min), an RNA polymerase inhibitor, we measured RNA abundance over time by RNA-seq. To extract position specific information within the RNA, we binned the reads every 300 nucleotides (nt) from the translation start site along the annotated transcript. A global picture of the data was obtained by averaging the respective 300 nt bins across all available transcripts (Figure 3.1A).

RNA abundance at the 5' end declines immediately after the addition of rifampicin (Figure 3.1B). The abundance at positions approaching the 3' end stays at

steady state for some time before declining. The amount of time the data in a bin spent at steady state is proportional to the distance of the bin from the 5' end (Gigure 3.1). This trend was not observed in the five-probe microarray data (Figure 3.2), which showed that abundance at the 3' positions was less abundant than steady-state at all time points. A likely explanation is that the trend was masked because the data was averaged over relative positions on different transcripts.

The two slopes in the RNA-seq data (Figure 3.1) likely reflect the existence of different processes. The line at y = 1 shows positions along the RNA where RNA synthesis and degradation rates remain at steady-state levels to produce steady-state RNA abundance. The positive slope shows where RNA is less abundant than at steady-state, showing positions that are experiencing only degradation. The 5' end is the first to show a decline in RNA abundance, and more positions downstream decline in abundance over time, reducing the number of positions that remains at steady-state abundance. Taken together, the data suggests that synthesis stops first at the 5', and the effect propagates along the transcripts towards the 3' end, resulting in a positive slope along the transcript. Before RNA synthesis is disrupted, RNA abundance remains at steady-state abundance. This interpretation is consistent with the fact that rifampicin stops RNA polymerases at initiation [17], leaving the other RNA polymerases to proceed on.

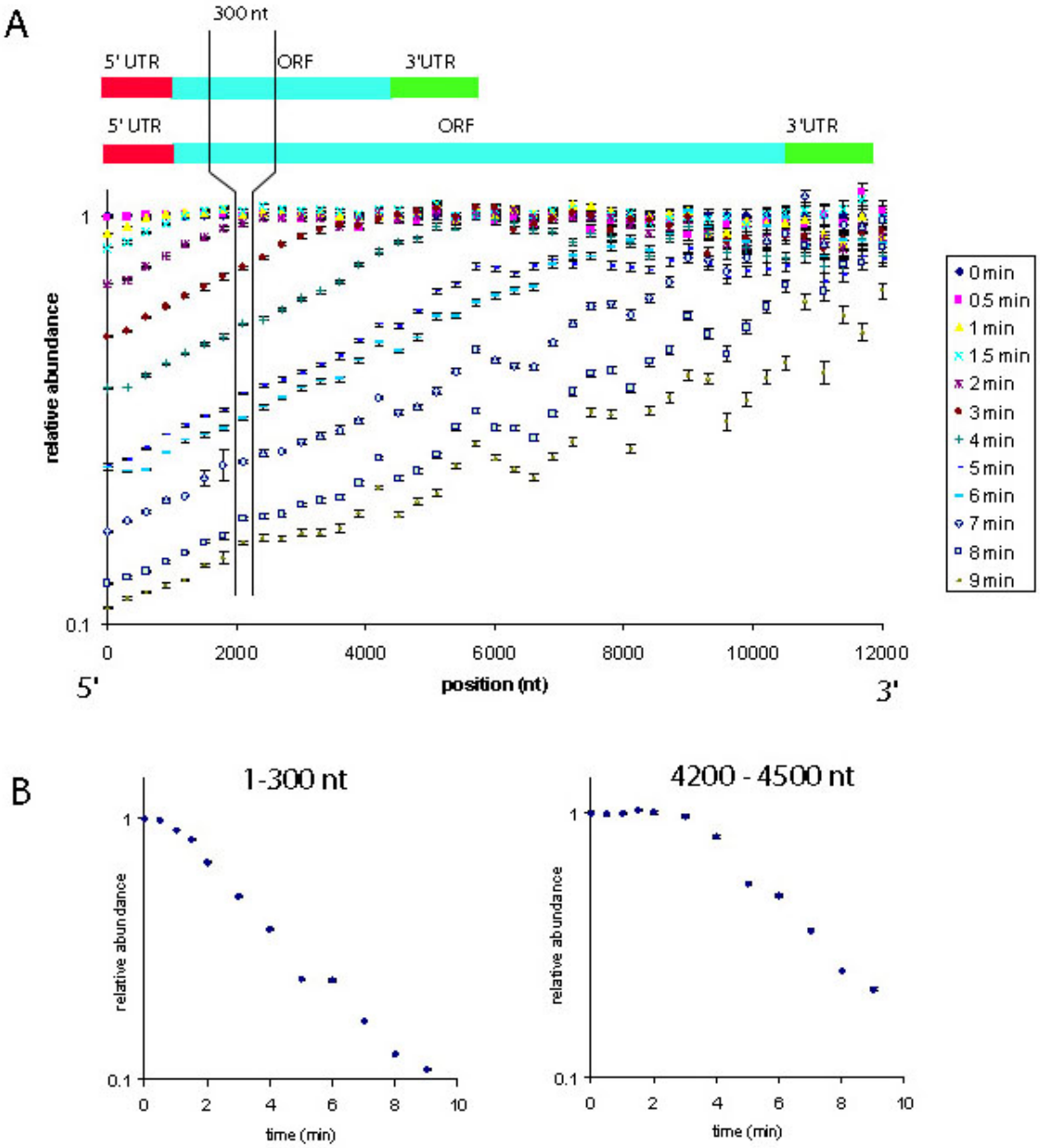


Figure 3.1 Global view of degradation measured by RNA-seq. To obtain position-specific data, the RNA-seq reads were binned every 300 nucleotides (nt) along each transcript. The number of reads (a readout for abundance) at each time point is normalized to the number of reads at time 0 min. The data from corresponding 300 nt bins from different transcripts were averaged. (A) The data is plotted by position on the transcript (x-axis). After the addition of rifampicin to stop RNA synthesis, RNA abundance decreases over time, reflecting degradation. The abundance of the 5' end decreases before the 3' end's, resulting in a positive slope. (B) The RNA-seq data is plotted over time for two specific bins, 1 – 300 nt, and 4200 – 4500 nt. For the first 300 nt bin, the action of degradation is apparent almost immediately after rifampicin was added. For the 4200 – 4500 nt bin, RNA abundance remains at steady-state levels for 2 – 3 min before declining.

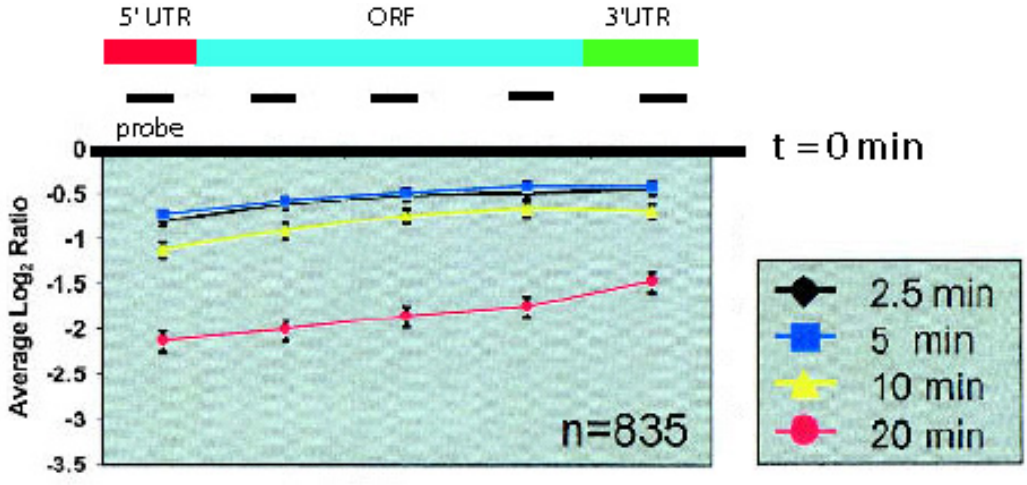


Figure 3.2 Global view of RNA degradation from the five-probe microarray. This figure is adapted from figure 1 in Selinger et al.[20]. The steady-state abundance (at t = 0 min) is marked. Five probes are designed to measure each RNA transcript, and each datapoint represents the average over all transcripts (n = 835) at the same probe position.

If it is true that residual RNA polymerase activity is responsible for the delay before synthesis stops at the 5' distal positions, stopping RNA synthesis by an elongation inhibitor should eliminate the apparent differences across the transcripts. Indeed, when RNA synthesis is stopped by elongation inhibitor streptolydigin, positions across the RNA show similar degradation behavior (Figure 3.3).

3.4 Extracting RNA lifetime

There are thus two separate parts in the RNA-seq data, each reflecting different processes. The first part is when the RNA stays at steady-state abundance, which reflects the action of residual polymerases and degradation. This part is fit by a line at y = 1 (Figure 3.4). Extraction of elongation data will be discussed in more detail later in this chapter (Section 3.7). The second part of the data shows only RNA degradation, and the data is usually fit to a single exponential function [1],

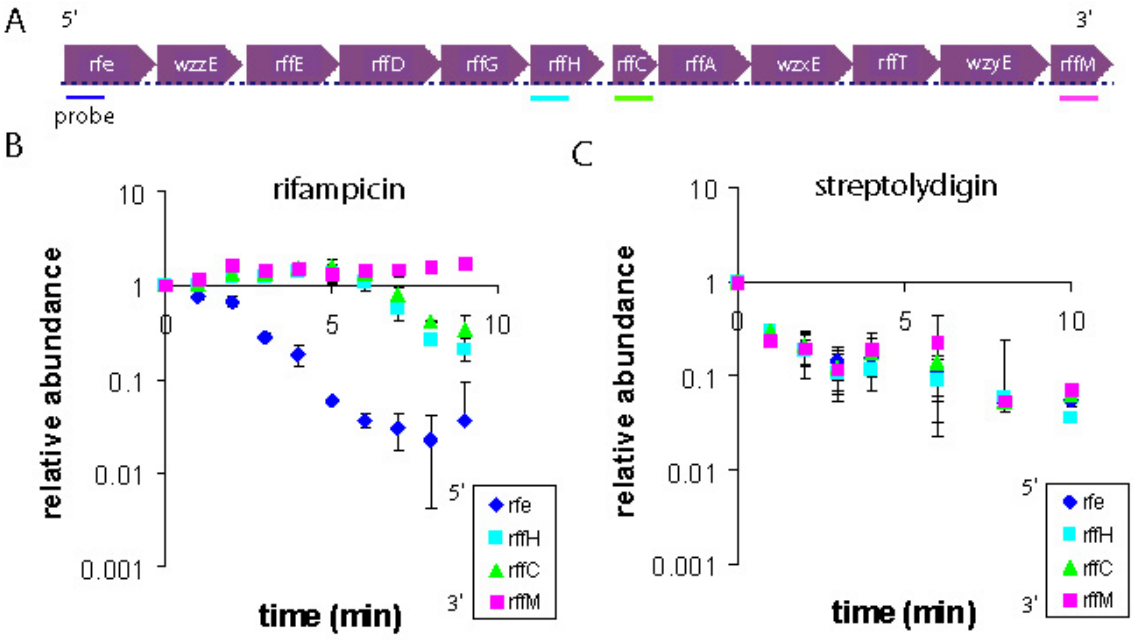


Figure 3.3 Residual polymerases cause downstream positions to stay at steady-state abundance after rifampicin addition. (A) Four quantitative PCR (qPCR) probe positions were picked for the ~12kb long transcript rfe-wzzE-rffEDGHCA-wzxE-rffT-wzyE-rffM. (B) RNA abundance at the four positions was measured over time, after the addition of rifampicin. Abundance at the 5'-most position (rfe) quickly declines, but the downstream probe positions remain at steady-state abundance for a few minutes. For the entire duration of the experiment, the rffM position never leaves steady-state. (C) RNA abundance at the four positions was measured over time, after the addition of streptolydigin. All four positions show immediate decline in RNA abundance.

$a(t) = a_o e^{-(t-d)/\tau}$, where a_o is the abundance of RNA at time 0 min (which is 1), to extract τ, the lifetime of the RNA. *d* accounts for the delay in stopping synthesis. Because the data was split into bins for analysis, the reported lifetime is for the 300 nt bin.

How do the lifetimes of 300 nt bins come together to give the lifetime of an RNA? Across a transcript, the lifetime of each 300 nt bin is very similar to that of the neighboring 300nt bins (Figure 3.4). Since different positions within a transcript have similar lifetimes, the lifetime of the transcript is defined as the average of the lifetime

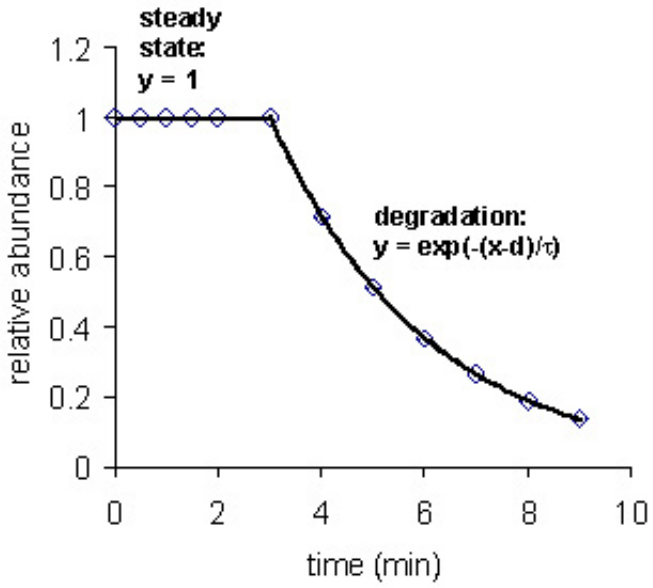


Figure 3.4 Extracting degradation rate information from the data. There are two parts to the data. In the first part, steady-state abundance is fit by y = 1. The remaining data points are fit by an exponential curve to extract τ, the RNA lifetime. d is the amount of time during which y = 1, and corrects for the delay before synthesis stops. The two fits are evaluated simultaneously to minimize the total residuals. The intersection of the two fits provides information about RNA elongation. It can occur between data points.

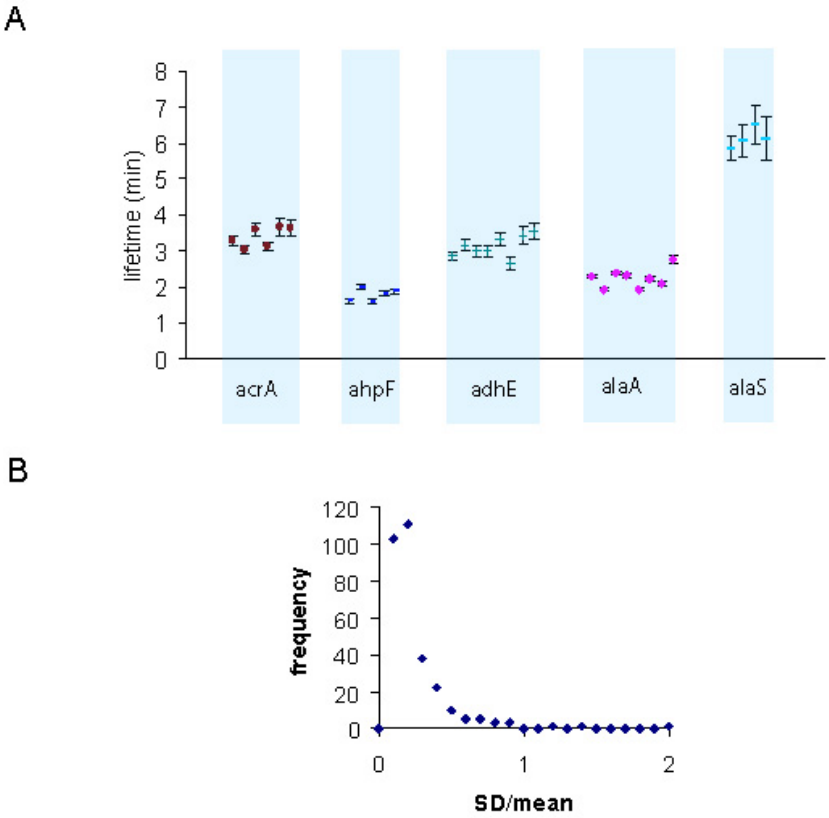


Figure 3.5 RNA lifetime is constant across a transcript, but different between transcripts. (A) The lifetime of each 300 nt bin is shown for five genes, acrA (pos: 483,650 <- 484,843), ahpF (pos: 638,976 -> 640,541), adhE (pos: 1,294,669 <- 1,297,344), alaA (pos: 2,405,583 -> 2,406,0800) and alas (pos: 2,817,403 <- 2820,033). The lifetimes within a transcript are similar, but each gene has a distinct lifetime from the other transcripts. (B) For 304 transcripts longer than 1800nt, the standard deviation (SD) of the lifetime of the first six bins was calculated, normalized by the mean lifetime, and plotted as a histogram. On average, the SD is 20% of the mean, suggesting that lifetimes are similar across a transcript.

of the bins within the transcript. Between transcripts, the lifetimes are distinct (Figure 3.4).

Our data is different from previous reports that the 5' ends of RNA were less stable than the 3' ends [20], and has implications on the models of RNA degradation. The observation that different parts of a transcript have different stabilities was used to

support the generality of 5' to 3' directionality of RNA degradation. The fact that 5' and 3' ends have similar stabilities is still compatible with 5' to 3' degradation (Section 3.10).

A prevailing model for RNA degradation posits a slow initiating step in degradation, followed by rapid degradation by other degradation enzymes [6]. This model predicts a longer lifetime on the 5' end than the 3' end, which is not compatible with our data. In order to make the model agree with the data, we need to include a protection mechanism to increase the lifetime of the 3' end. This could be protection by a translating ribosome, which prevents rapid degradation of the 3' end. Alternatively, one of the degradation enzymes acting after the initiating enzyme needs to move more slowly, at the rates similar to that of RNA polymerase.

3.5 RNAs are less stable than previously thought

We were able to extract lifetimes for 956 transcripts. Consistent with literature, we found that RNA lifetimes are typically less than 10 minutes. However, the peak of our RNA lifetimes was around 2-3 minutes, in contrast to the peaks at 4-7 minutes for the previously published measurements (Figure 3.6). Besides the different experimental conditions, the shorter lifetimes are also the result of our higher resolution data, which provided impetus for a new data analysis scheme that corrects for residual RNA polymerases.

The microarray dataset had five time points at 0, 2, 4, 6 and 8 min, and performed a least squares linear fit of the log of the abundance data [3]. The accepted fits had $R^2 > 0.7$. We mimic the data analysis of the microarray paper with our 12-point dataset. Using the corresponding five points, the extracted lifetime is 3.5 min (Figure 3.7A). Although the R^2 value is high, the value of a_o is 1.4

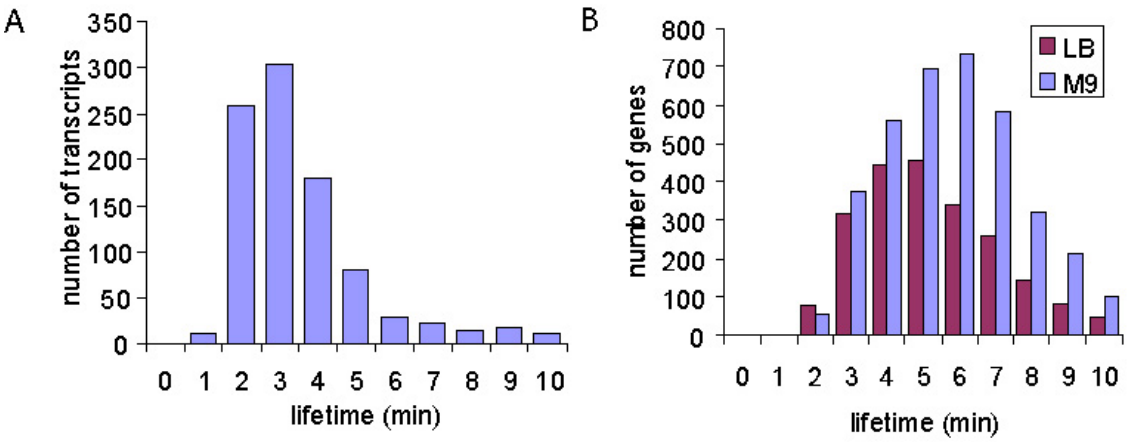


Figure 3.6 RNAs are less stable than previously thought. (A) The distribution of lifetimes from cells grown in LB at 30C and measured by RNA-seq peaks at 2-3 min. The lifetimes reported were extracted while taking elongating polymerases into account. (B) Figure adapted from figure 1 of Bernstein et al. [3]. The peak of RNA lifetime distribution is at 4-5 min for cells grown in LB, and 5-6 min for cells grown in M9. The elongating polymerases were not taken into account in this lifetime measurement.

instead of the expected 1. It is not possible to judge if the 0 or 2 min time points deviate from the curve because of experimental noise or a delay before RNA synthesis stops.

Viewing the twelve-point RNA-seq dataset for cyoABCDE, it becomes obvious that RNA synthesis took over 2 min to stop, thus time points before 2 min are unsuitable for fitting to a decay curve (Figure 3.7B). The global fitting by y = 1 and the exponential curve results in an adjustment of 2.5 min for the time it took for synthesis to stop, and gives a lifetime of 2.4 min. The a_o is close to 1.

3.6 Exceptions to single exponential decay

Even after correcting for the delay in stopping RNA synthesis, a small fraction of the data did not fit well (low R^2, ao ≠ 1) to a single exponential decay. The poor fitting is caused by the lack of apparent degradation at the later time points (Figures 3.8A and 3.8B). The earlier time points may fit well to a single exponential curve if the later time

points are excluded, suggesting that RNA degradation was initially happening in a typical manner.

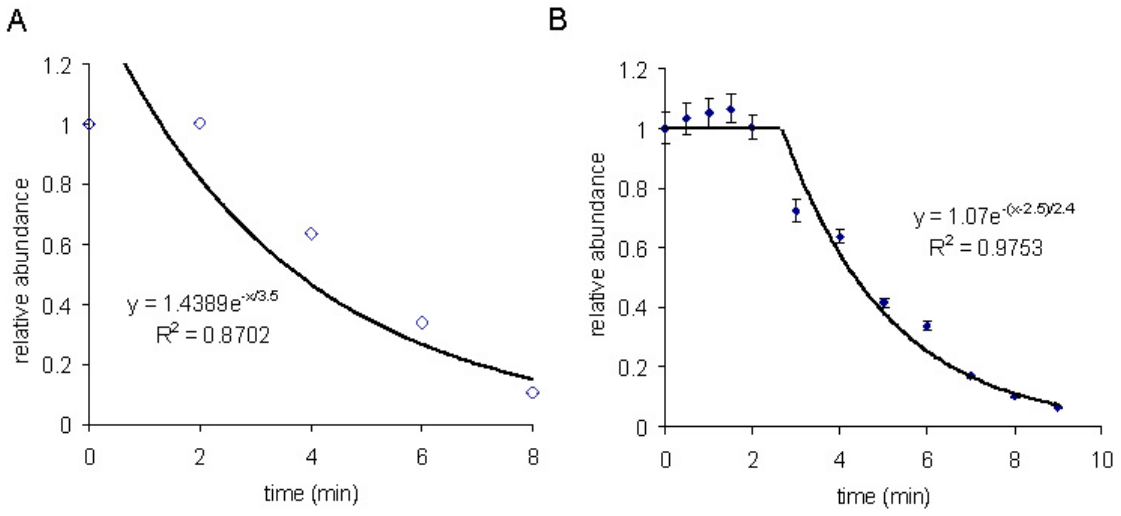


Figure 3.7 Incorrect fitting of data results in longer RNA lifetime. RNA-seq data from the 1 – 300 nt bin of the cyoABCDE transcript is shown. The data is scaled to the 0 min time point. (B) The 0, 2, 4, 6, 8 min time points corresponding to the microarray experiment are shown. Fitting all five time points to an exponential results in a lifetime of 3.5 min ($R^2 = 0.8702$). (A) The entire dataset of twelve time points are shown. Relative abundance stays at 1 for 2.5 minutes, before it starts to decline, indicating that synthesis has stopped. The exponential fit gives a lifetime of 2.4 min.

We confirmed that the apparent lack of degradation is not caused by insufficient data collection by ensuring each 300 nt bin had at least 20 reads at the last time point. We also observed similar behavior over time by qPCR (Figure 3.8D).

The lack of degradation is not a system-wide effect, apparent from the fact that 95% of RNAs exhibit single exponential decay during the same measurement period. The apparent amount of un-degraded RNA varies (Figure 3.8A – B), suggesting that the effect is transcript-specific. There are many possible mechanisms

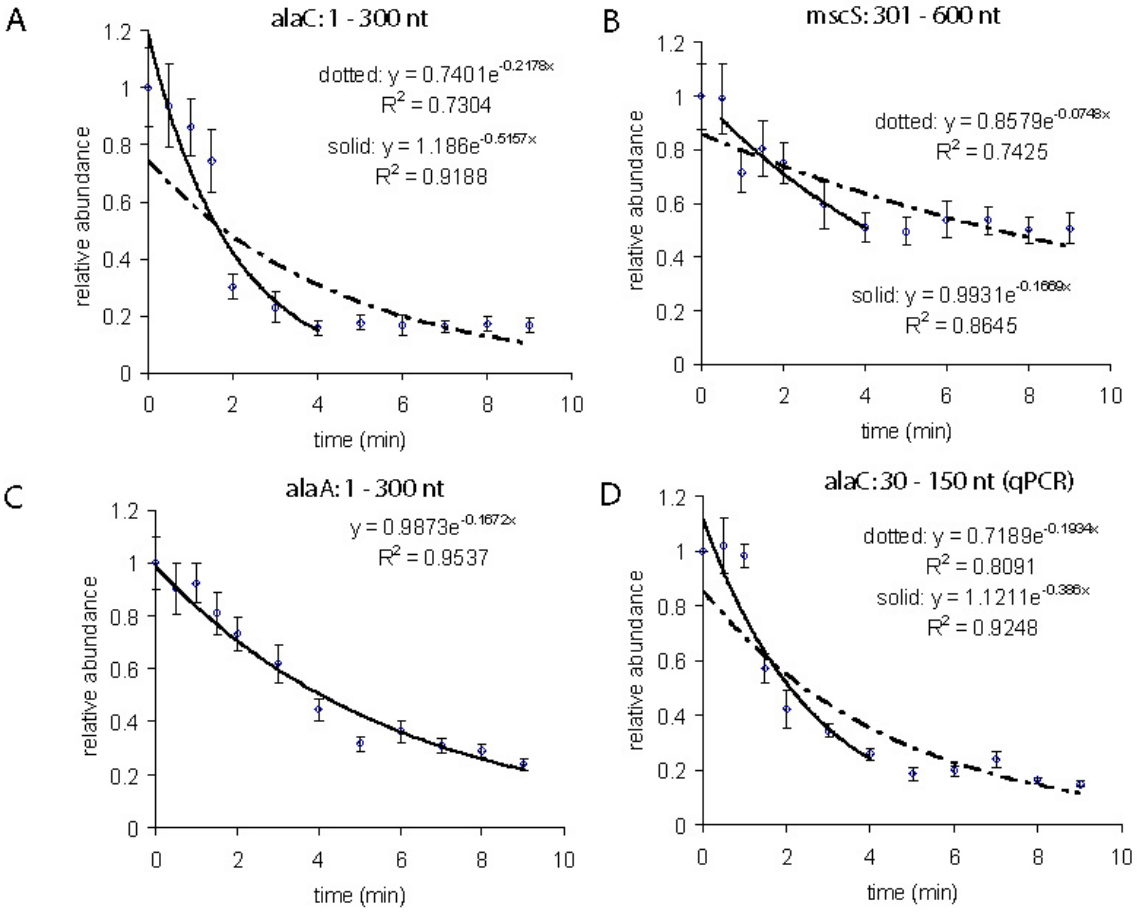


Figure 3.8 Evaluating fits to a single exponential RNA decay. (A) RNA-seq data from the 1 – 300 nt bin of alaC is shown. RNA abundance at time points after 4 min stays at ~15%. The data does not fit a single exponential curve well (dotted line). The initial decay (0 – 4 min) fits a single exponential curve. (B) RNA-seq data from the 301 – 600 nt bin of mscS is shown. RNA abundance at time points after 4 min stays at ~50%. The dotted curve is the result of fitting all data points to a single exponential, while the solid curve is the result of fitting the data from 0.5 to 4 min to a single exponential. (C) RNA-seq data from the 1 – 300 nt bin of alaA is shown. The data fits well to a single exponential curve. (D) qPCR data with primers to the 30 – 150 nt fragment of the alaC RNA. RNA abundance after 5 min stays at ~15%, similar to (A).

to explain the behavior of this sub-population of transcripts, including anti-sense RNA, and secondary structure and rifampicin-related artifacts.

3.7 Extracting RNA elongation rates

Conventionally, elongation rate measurements have been done on a limited basis in two different ways. The first method uses rifampicin and radioactive labeling of

nascent RNA [19]. Rifampicin is added to a culture and at intervals after, aliquots are removed and exposed to 3H-uridine for a defined period of time. The amount of 3H-uridine incorporated is proportional to the number of RNA polymerases remaining on the DNA. By this method, the elongation rates of the trp operon [19] and also ribosomal RNA [7] were measured in *E. coli* B/r.

The second method uses isopropyl-b-D-thiogalactoside (IPTG). The DNA encoding the gene of interest is cloned behind a lac promoter on a plasmid, and the abundance of RNA is measured at different probe positions on the gene over time after the addition of IPTG [21]. This method is more commonly used in recent papers [8, 22]. Researchers have also switched to using the *E. coli* K-12 strain in experiments.

Both methods have only been used to measure the elongation rate on one transcript at a time. For the IPTG method, only one transcript can be measured at a time because it is cloned onto a plasmid. The first method using rifampicin is actually very similar to our protocol, and although laborious, can be used to measure the elongation rates on multiple transcripts. However, the method is not suitable for measuring RNA elongation rates genome-wide.

A point of concern about the IPTG protocol is that the measured elongation rate has been shown to be dependent on the amount of IPTG used [8]. Thus while this method is suitable for comparing the elongation rate of different transcripts under the same conditions, it does not give the actual elongation rate from the native promoter in the cell.

In section 3.4, we discussed fitting of the data to two curves, $y = 1$, and $y = a_o e^{-(x-d)/t}$, where *d* is the delay before RNA synthesis stops. *d* is also the amount of time the last RNA polymerase in the sample takes to reach a particular bin of a transcript, which we

call the polymerase passage time. Each bin on a transcript has a RNA lifetime, and a polymerase passage time (Figure 3.9A – B). If RNA polymerases travel at approximately constant rates over a transcript, by tracking the movement of the last RNA polymerase over time, we can extract the average RNA elongation rate (Figure 3.8).

We imposed a few criteria to ensure the accuracy of the elongation rates measured. Firstly, we required at least six data points in a fit, limiting the transcript length to a minimum of 1800 nt (at least six 300 nt bins). In addition, we only fitted for a constant velocity (linear fitting) because the resolution of the data was not high enough to determine otherwise.

We were able to extract the elongation rates of 106 transcripts from our data (Figure 3.9). This is the first time multiple elongation rates have been measured simultaneously from native promoters. The distribution of the elongation rates is asymmetric, with a heavy right tail. The average elongation rate for the 106 transcripts was 23.5 nt/s, which is where the distribution peaks.

For many years, the elongation rate at was expected to be around 50nt/s for mRNA and 75 – 100 nt/s for ribosomal RNA based on the early work on *E. coli* B/r measured with rifampicin, and the later work measured in *E. coli* K-12 with IPTG. Even adjusting for the different temperatures the experiments used, our measured distribution of mRNA elongation rates suggests that the early work measured

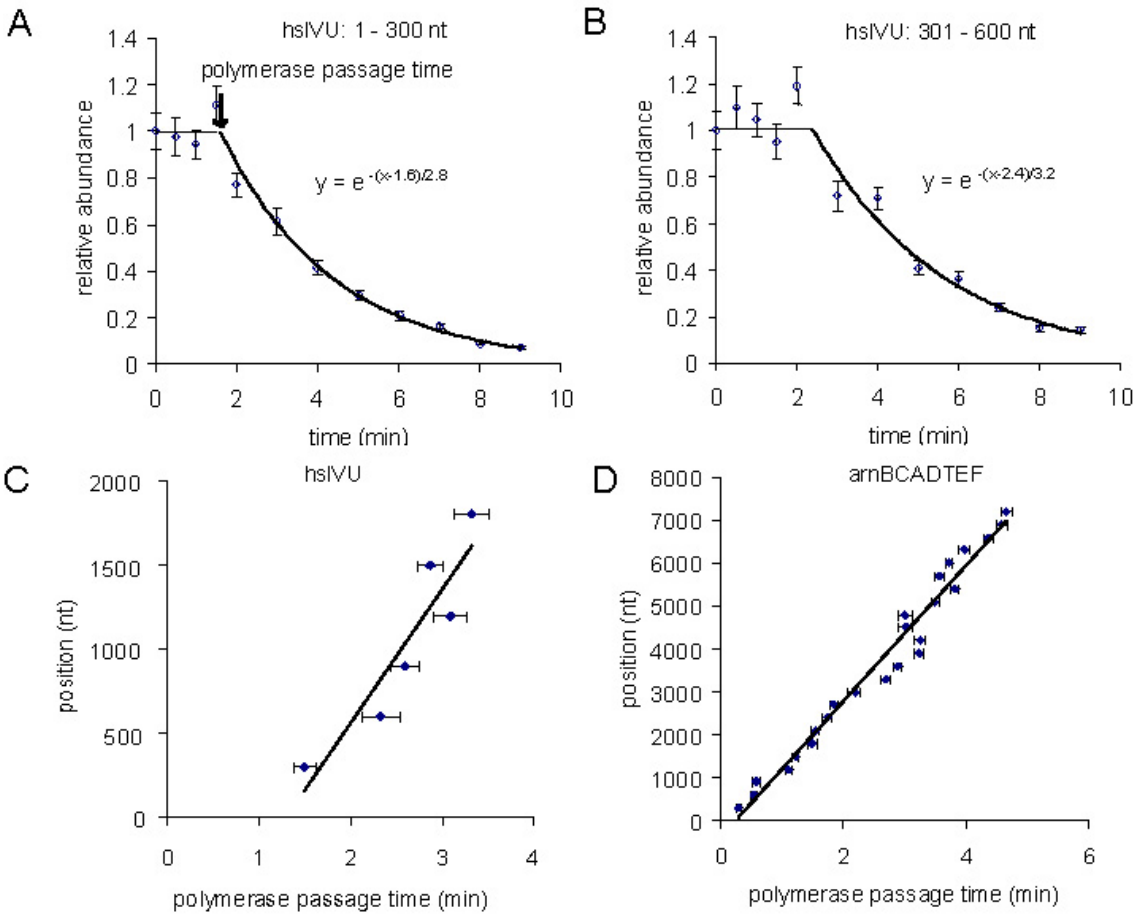


Figure 3.9 Extracting RNA elongation rate from RNA-seq data. (A) Data from the 1 – 300 nt bin of hslVU is shown. It takes 1.6 min before RNA synthesis stops. This intersection between the two fitting curves is the polymerase passage time – the time it took for the last RNA polymerase to reach the bin. (B) Data from the 301 – 600nt bin of hslVU is shown. The polymerase passage time is 2.4 min. (C) The polymerase passage time and bin position are plotted for hslVU, and the elongation rate is given by the slope of the linear regression. The average last RNA polymerase proceeds at an approximately constant velocity of 14.6 nt/s. (D) The polymerase passage time and bin position are plotted for arnBCADTEF with the linear fit. The average last RNA polymerase also proceeds at an approximately constant velocity of 26.5 nt/s.

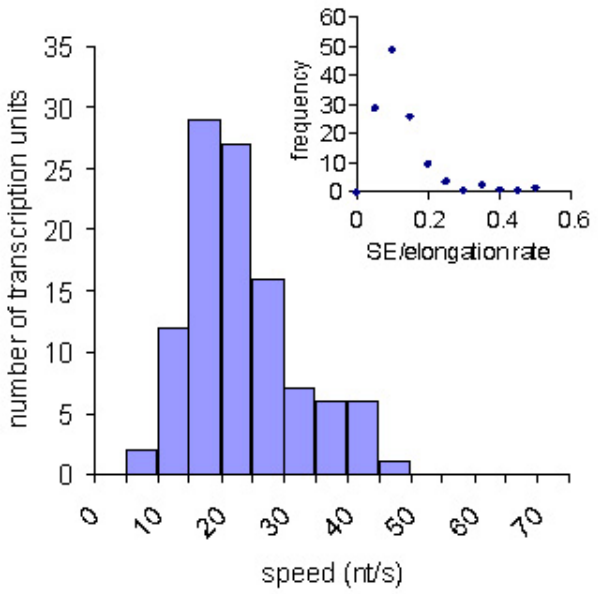


Figure 3.10 Histogram of elongation speeds for 106 transcripts. The average elongation speed for 106 transcripts is 23.5 nt/s. The distribution is slightly skewed with a long right tail. (Inset) The distribution of the uncertainty of the measurement (standard error) is shown. The uncertainty averages about 10%.

transcripts that were transcribed faster than average. We will discuss the significance of this finding in context with the other parameters we have measured in this experiment in section 3.11.

3.8 Calculating steady-state RNA abundance and synthesis rates.

Besides RNA lifetime and elongation rate, we also have information about RNA abundance at steady state in the time = 0 min data point. We added non-*E. coli* RNA to the sample at amounts proportional to the number of cells. Using the relationship between the number of non-*E. coli* spike-in RNA added and the number of reads, we can convert the reads to copy numbers per cell [16] (Figure 3.11).

We find that RNA abundance spans six orders of magnitude (Figure 3.12). Transcripts containing essential genes are expressed at a higher level, above three copies per cell, and are more narrowly distributed. This is reminiscent of the

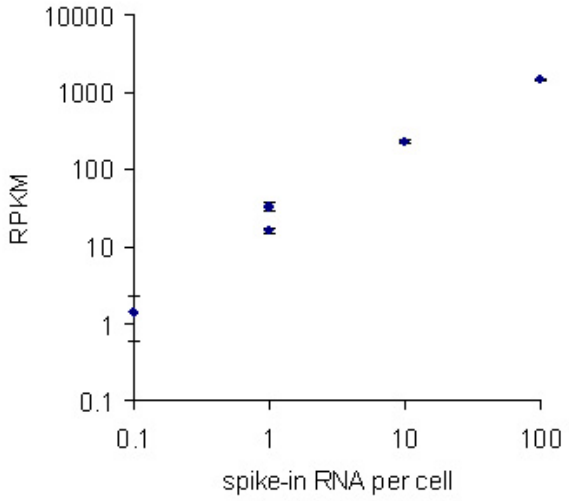


Figure 3.11 RNA-seq calibration curve. Six spike-in RNAs of varying lengths were added to the samples prior to RNA purification at different proportions to the number of cells. The number of reads was converted to reads per KB per million reads (RPKM) to adjust for transcript length [16]. The number of spike-in RNA per cell and RPKM are proportional ($R^2 = 0.997$).

abundance of essential proteins, which are more highly expressed than non-essential proteins at above ten copies per cell (Chapter 2). Because we cannot distinguish between the different RNA isoforms, direct counting of all possible transcripts double-counts of the number of RNAs in the cell and produces a total RNA count that is an order of magnitude larger than other estimates [4]. With better annotation, these estimates should come closer to agreement.

RNA abundance, synthesis rate, and degradation rate are related in the following manner,

$$[\text{RNA}] = \frac{S}{k_d} = S \,.\, \tau$$

where S is the RNA synthesis rate, k_d is the RNA degradation rate, and τ is the RNA lifetime. k_d is the inverse of lifetime, τ.

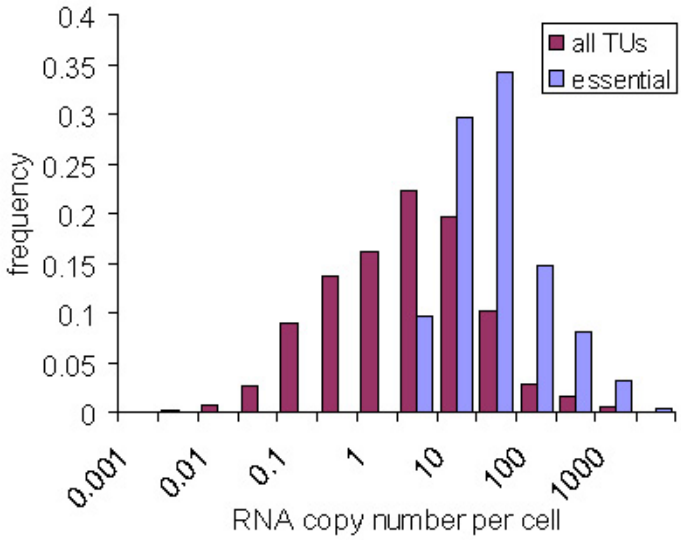


Figure 3.12 Distribution of RNA copy numbers per cell. The copy number of 2746 transcripts is histogramed, of which 257 are transcripts containing essential genes. The y-axis is gives the fraction of the population (all transcripts, or transcripts with essential genes) that exist in the respective copy numbers per cell. RNA copy numbers span six orders of magnitude. Essential transcripts tend to be highly expressed, at above three copies per cell.

Using the RNA copy number and lifetimes, we calculated RNA synthesis rates. Synthesis rates span four orders of magnitude (Figure 3.13A), and correlates well with RNA abundance, accounting for most of the variation in RNA abundance (Figure 3.13B). On the other hand, RNA degradation rates are not correlated with RNA abundance (Figure 3.13C, $R^2 = 0.00$). This seems different from the published claim that RNA abundance and degradation rates are correlated [3]. Two datasets for cells grown in M9 or LB media were collected for the Bernstein paper, of which only the sample grown in M9 had the correlation and was presented and the LB data was ignored. Our results are therefore consistent.

Lastly and unexpectedly, we found a small correlation between RNA synthesis and degradation rates when synthesis rates are low (Figure 3.13D, $R^2 = 0.37$). This correlation could be due to the spatial arrangement of genomic DNA, and its resulting accessibility to RNA polymerases and RNases, but needs to be substantiated by experiments.

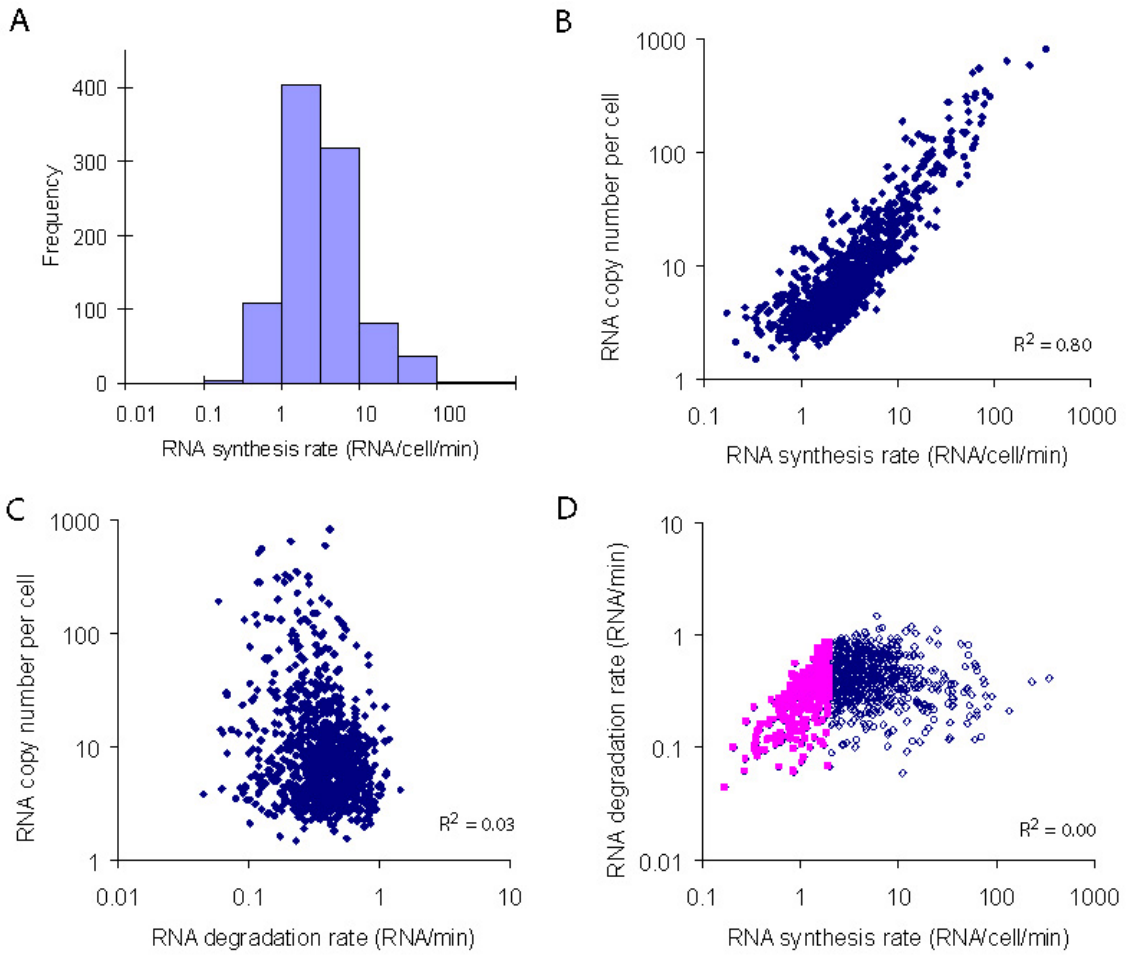


Figure 3.13 Correlations between RNA copy number, RNA synthesis and degradation rates. RNA synthesis rates were calculated for 956 transcripts using RNA abundance and degradation rates. (A) The distribution of RNA synthesis rates ranges almost four orders of magnitude, peaking around 1-3 transcripts/cell/min. (B) RNA synthesis rates and RNA copy number per cell are correlated ($R^2 = 0.80$). Most of the differences in RNA copy numbers can be explained by the differences in RNA synthesis rates. (C) RNA abundance and degradation rates are not correlated ($R^2 = 0.03$). (D) RNA degradation and synthesis rate are not correlated for the entire dataset ($R^2 = 0.00$, n = 956). However at low rates of synthesis (<2 RNA/cell/min, magenta, solid), there is a small correlation between degradation and synthesis rates ($R^2 = 0.37$, n = 349).

3.9 Determinants of elongation rate

It was previously observed that the RNA polymerases can act cooperatively to promote faster elongation rates [8]. As a result, we expected to find a correlation between RNA synthesis rates and elongation rates. However, there was no correlation (Figure 3.14A, $R^2 = 0.01$).

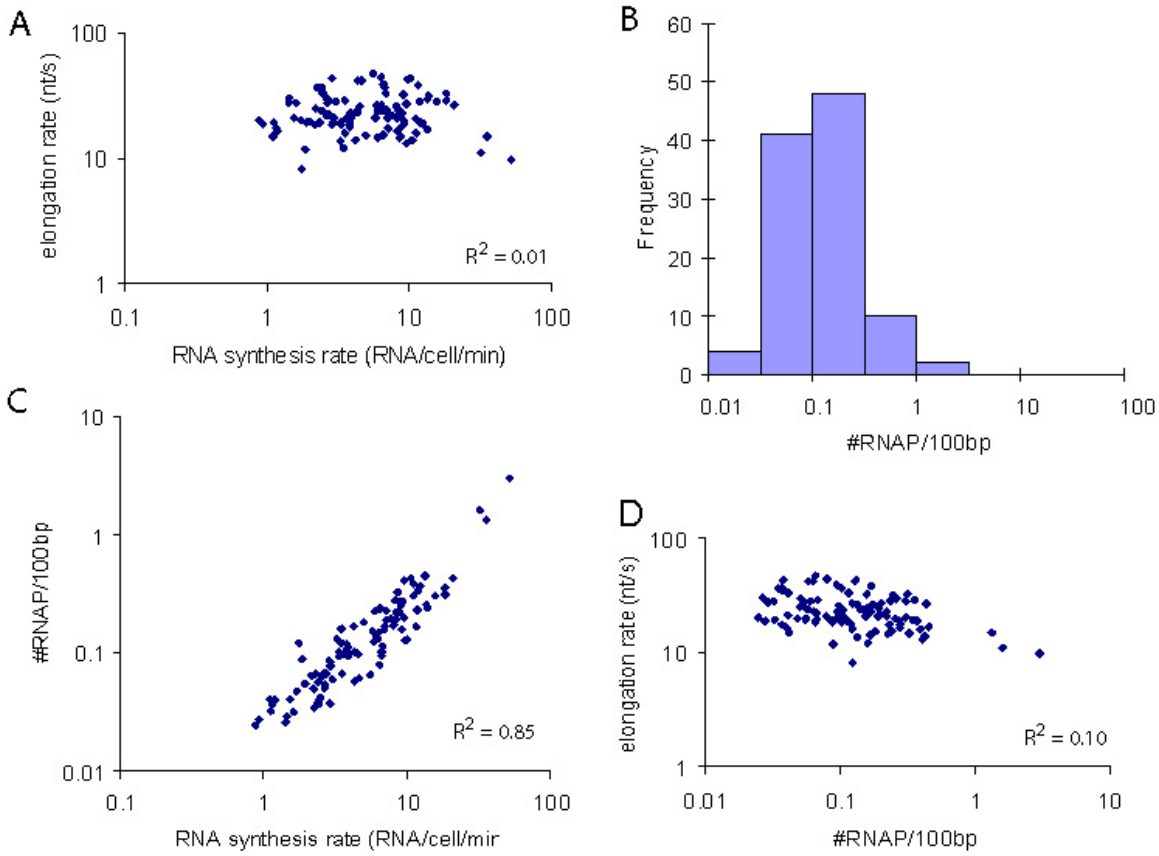


Figure 3.14 Correlations with elongation rates. RNA elongation rates were measured for 106 transcription units. (A) Elongation rate correlates poorly with RNA synthesis rate ($R^2 = 0.01$). (B) The distribution of RNAP density is shown. Most RNA polymerases (RNAP) are on the average 300 – 3000 bp away from each other. (C) Average RNAP density was calculated assuming three copies of genomic DNA per cell. RNA synthesis rate and average RNAP density are correlated ($R^2 = 0.85$). (D) Elongation rate correlates poorly with average RNAP density ($R^2 = 0.10$).

We reasoned that RNA synthesis rate does not accurately reflect the proximity between RNA polymerases, because it does not take into account polymerase speed

(elongation rate) and the length of the transcript, and thus may show a weaker correlation with elongation rates. The relationship between RNA synthesis rate, *S*, and the density of RNA polymerases is as follow:

$$S = k_s \cdot D = \frac{N \cdot v}{L}$$

where k_s is the rate constant, *D* is the number of genomic DNA for the transcript, *N* is the number of RNA polymerases, *L* is the length of the transcript, and υ is the elongation rate. Thus the average density of RNA polymerases is

$$\frac{N}{L \cdot D} = \frac{k_s}{v}$$

We currently assume $D = 3$, although the number of copies of genomic DNA varies along the genome [4].

We obtained the distribution of RNA polymerase densities for 106 transcripts (Figure 3.14B). The distance between two RNA polymerases is typically on the order of 300 - 3000bp. Moreover, there is only one RNA polymerase on some of the transcription units. Because our measurement is biased towards longer transcripts that are also more abundant, extrapolating from these 106 transcripts to estimate the number of active RNA polymerases in the cell would result in a number that is an order of magnitude larger than other measurements [4].

There is a strong correlation between RNA synthesis rate and the average density of RNA polymerases (Figure 3.14C, $R^2 = 0.85$). The RNA synthesis rate represents the loading rate of the polymerases onto the DNA template, and thus is expected to explain the density of RNA polymerases. We expect the correlation to be slightly weaker when *D* varies.

The correlation of elongation rate with average RNA polymerase density is stronger than the correlation with RNA synthesis rate, but is still small (Figure 3.14D, R^2 = 0.10). The small correlation is expected given that RNA polymerases are too far to physically interact cooperatively (Figure 3.14B). The improvement of the correlation between elongation rate and average polymerase density compared to with RNA synthesis rate then becomes surprising because there is no physical cooperation. If this negative correlation is true, it is suggestive of a memory effect on the DNA.

It is known that the speed of RNA polymerase is affected by interaction with protein co-factors and RNA sequences [2, 18]. While these interactions are not well studied on a genome-wide level, future findings may allow researchers to understand how they affect RNA elongation rates.

3.10 Organization of RNA polymerases and ribosomes in a bacterial cell

The density of RNA polymerases on the genome can be directly observed by electron microscopy of a chromatin spread [12]. In a bacterial system, the translating ribosomes can also be observed on the RNA [13] (Figure 3.15). Due to the lack of identifiable markers, chromatin spreads are limited to studying highly-expressed ribosomal RNA operons and the neighboring genes. In section 3.8, we were able to calculate the density of RNA polymerase genome-wide using RNA synthesis rates and elongation rates. A similar calculation can be performed to obtain the density of ribosomes

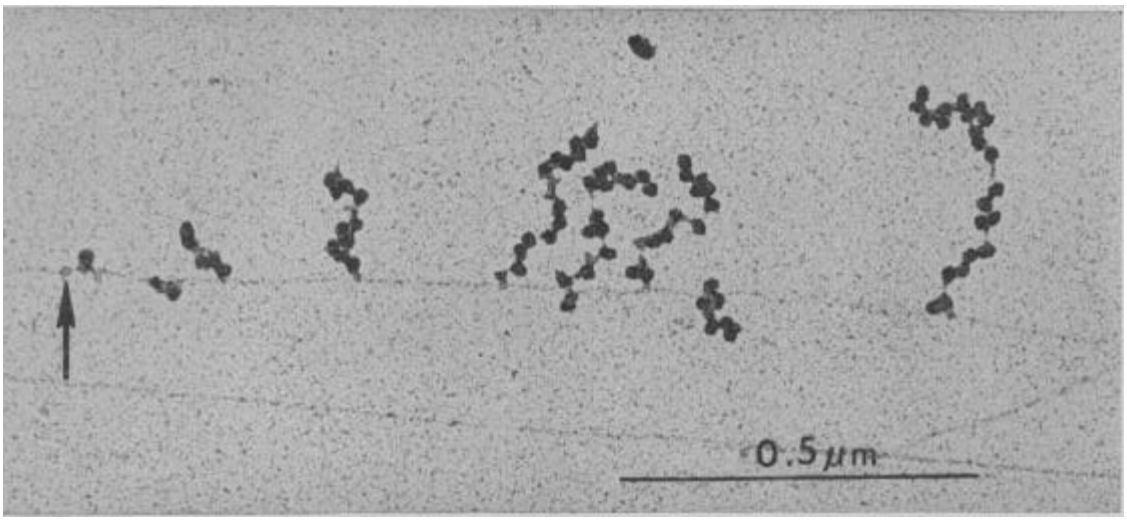


Figure 3.15 Electron microscope image of translating ribosomes, RNA polymerases and nascent RNA on the chromosome of E. coli. This image is adapted from Miller at al. (1970). An RNA polymerase presumed to be on the promoter is marked by the arrow, and the ribosomes are the larger and denser objects. The segment of chromosome shown could not be identified.

A relationship analogous to that of RNA synthesis rate and density of RNA polymerase can be constructed for protein synthesis rate and ribosome density

$$\frac{N_r}{L_p \cdot RNA} = \frac{k_p}{v_r}$$

where N_r is the number of ribosomes on an RNA, L_p is the length of the protein-coding RNA in nt, k_p is the protein synthesis rate (protein/RNA/s), v_r is the elongation rate of the ribosome in nt/s, and *RNA* is the steady-state copy number of RNA per cell.

While protein synthesis rates and ribosome speeds are not available from our RNA-seq data, the former is available from the YFP library data (Chapter 2) albeit obtained under different experimental conditions. To use the existing data, we assume an elongation rate of 20nt/s, approximately the average speed of RNA polymerases at 30C (our measurement), for the ribosomes. If ribosomes are additionally affected by the availability of amino acids, the elongation rate will be slower, resulting in a higher density of ribosomes. We assumed one elongation rate for all proteins for simplicity.

We were able to obtain the protein synthesis rate and RNA copy number for 498 genes. The resulting distribution of ribosome density has a long left tail (Figure 3.16), a consequence of the single-molecule sensitivity of the experiment. This long left tail suggests that some RNA never get translated (0.001 ribosome/100bp is a distance of 100,000 bp between ribosomes). The peak of the distribution is between 0.1-1 ribosome/100bp, which is slightly higher than the density of RNA polymerase. This difference in density will become more pronounced if the ribosome elongation rate is slower.

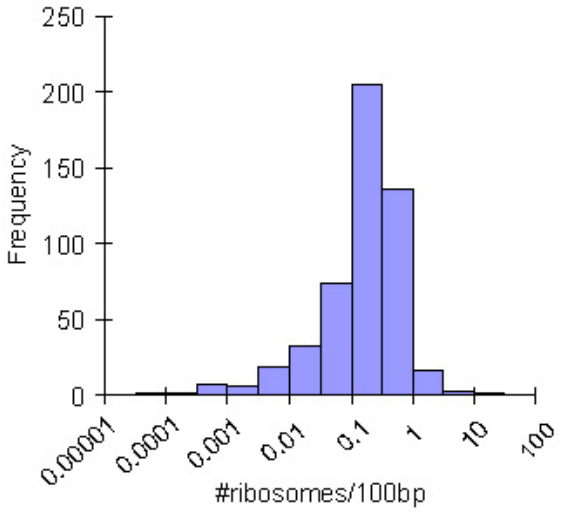


Figure 3.16 Distribution of ribosome density. The average number of ribosomes on 100bp of RNA is calculated for 498 genes using data from the YFP library (Chapter 2) and assuming a ribosome elongation rate of 20nt/s. The peak of the distribution is at 0.1-1 ribosomes/100bp, suggesting a gap of 100-1000bp between ribosomes.

We find that the relative abundance of RNA in cells grown in M9 or LB media is similar ($R^2 = 0.55$, Figure 3.17A). Other researchers have found that relative protein abundance in cells grown in rich or poor media are correlated ($R^2 = 0.63$) [11]. Although the RNA polymerase and ribosome densities were obtained under different conditions, their relative densities should be proportional because RNA and protein abundances are

proportional. We can thus ask whether there are patterns in the distribution of RNA polymerases and ribosomes.

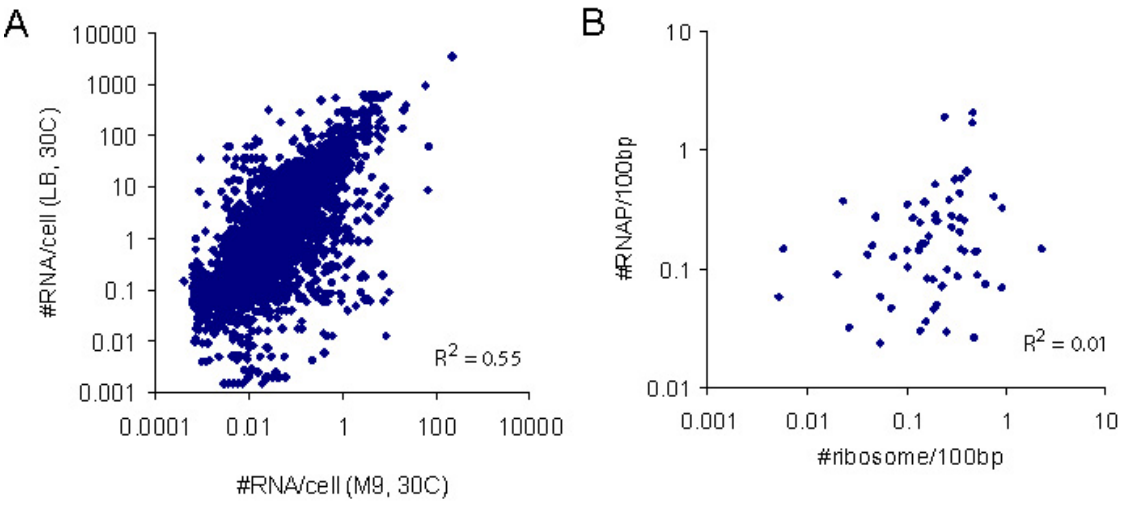


Figure 3.17 Strategies of gene expression. (A) RNA abundance in two different K-12 sub strains grown under different conditions are correlated ($R^2 = 0.55$). (B) Each dot represents the RNA polymerase and ribosome densities for a gene. Data was available for 61 genes. RNA polymerase and ribosome densities are not correlated in these genes ($R^2 = 0.01$).

We examined the relative density of RNA polymerases and ribosomes for 61 genes. Our dataset was limited by the number of genes that had both an elongation rate, and a protein synthesis rate. We find no correlation between RNA polymerase and ribosome densities ($R^2 = 0.01$, Figure 3.17B). This means that the cell has different ways of controlling protein production.

3.11 Co-transcriptional degradation of RNA

How does the lifetime of an RNA compare to the amount of time it takes to synthesize the full length RNA? We can calculate the time needed to make an RNA, the synthesis time, using the RNA elongation rate and length.

For transcripts longer than 1800 nt, the distribution of synthesis times is similar to the distribution of RNA lifetimes, peaking at around 2 min (Figure 3.18A). This means that RNA synthesis and degradation occur on similar timescales. RNA lifetime and synthesis time are not correlated (Figure 3.18B).

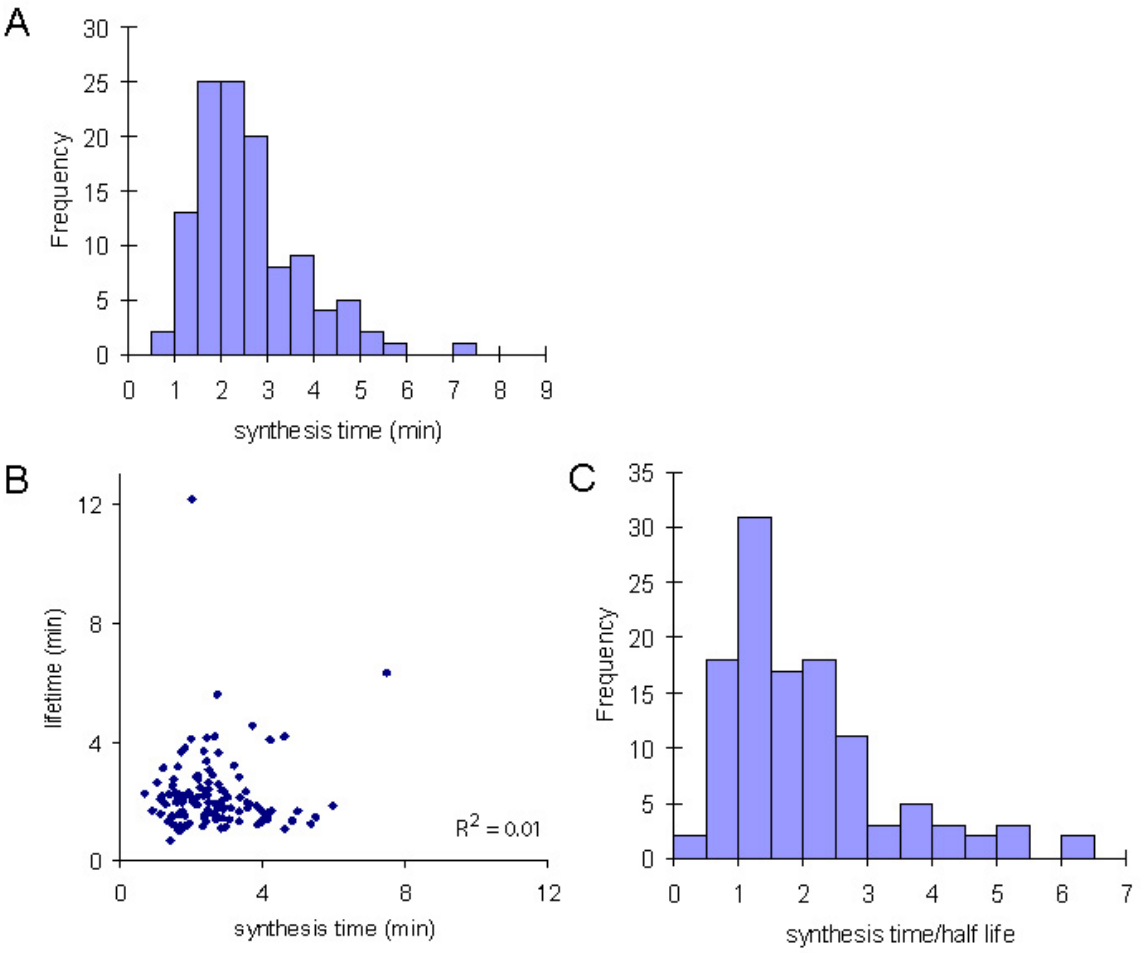


Figure 3.18 Comparing RNA synthesis time and RNA lifetime. (A) Synthesis time was calculated for 106 RNAs using the transcript length and elongation rate. The distribution peaks at around 2 min, similar to RNA lifetime. (B) RNA lifetime and synthesis rate are not correlated (R^2 = 0.00). (C) The ratio of the synthesis time to the RNA half life of the first 300nt bin is calculated, and the distribution is shown. Most RNAs have a synthesis time about 1 – 3 times longer than the half life.

For a more intuitive understanding of RNA stability, we convert the lifetime (τ) to half life ($t_{1/2}$)

$$t_{1/2} = \tau \ln 2$$

The half life is shorter than the lifetime, and represents the amount of time it takes for a population to decay to half the original amount. We took the ratio of the synthesis time to the half life of the 5' end (Figure 3.18C). This ratio gives the number of half lives the 5' end has undergone before the 3' end of the transcript is synthesized. While most RNAs have a synthesis time that is similar to the half life, some RNAs have a synthesis time that is at least four times longer than the half life.

To better understand what it means for the synthesis time to be longer than the half life, we describe the following framework. We assume that RNA can only be degraded after the entire transcript has been synthesized. This means that the 5' end will survive for at least as long as it takes for the transcript to be synthesized. We next assume that degradation happens very quickly. In this situation, the relative abundance of the 5' end is linearly proportional to the amount of time since the addition of rifampicin (Figure 3.19, open squares). If degradation is slow or delayed, the relative abundance over time is higher, and the resulting curve should shift right. On the other hand, if degradation of the 5' end happens before the transcript is synthesized, the curve is shifted left.

In the case where synthesis time is much longer than the half life, the measured relative abundance of the 5' end is smaller than expected by the model, suggesting that the 5' end is being degraded co-transcriptionally (Figure 3.19).

In general, longer transcripts have a higher chance of having a much longer synthesis time relative to half life, and are thus likely to be co-transcriptionally degraded. Indeed, the two known cases of co-transcriptional degradation in *E. coli*, lacZ [5] and the trp operon [14, 15] are relatively long transcripts.

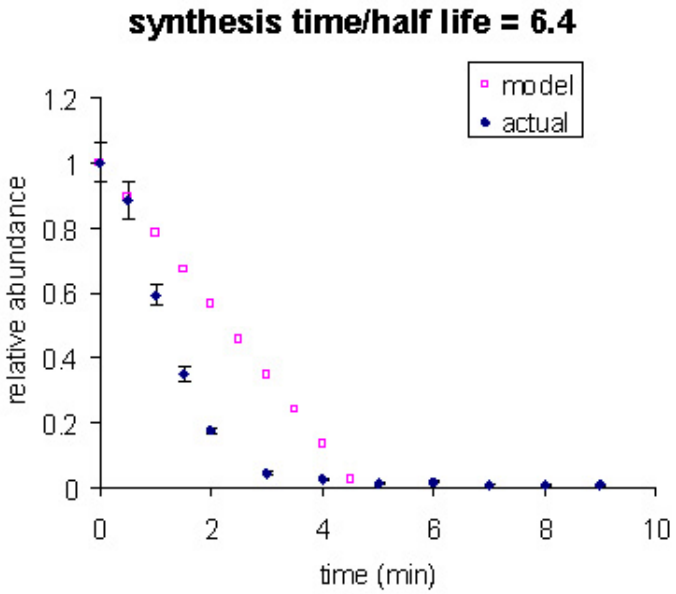


Figure 3.19 Evidence for co-transcriptional degradation. The relative abundance of the 1-300 nt bin is plotted for arnBCADTEF. The open squares predict the relative abundance over time using a model of no co-transcriptional degradation and very fast degradation of the 5' end upon completion of the transcript. The closed diamonds show the measured relative abundance. The measured relative abundance of the transcript is smaller than predicted by the model, strongly suggesting that this transcript is co-transcriptionally degraded.

A recent paper reported that RNAs tend to be located near their sites of synthesis on DNA [10] in *Caulobacter cresentus* and *E. coli*. The authors heavily relied on a 120 lacO repeat on the end of the RNA as an RNA fluorescence *in situ* hybridization (FISH) target to image the RNA in their experiment. This 120 lacO repeat is approximately 4.5 kb. If the elongation rates under their experimental conditions are similarly slow, the RNAs being observed are likely co-transcriptionally degraded. In this case, one would expect the RNA signal to exist mainly near the site of synthesis. The experiment could have been better designed if the researchers did not start with an assumption that elongation rates are fast, and thus RNA synthesis times are negligible.

3.12 Summary

In order to better understand RNA regulation, we measured steady-state RNA abundance and RNA degradation rates genome-wide using RNA-seq. Because RNA-seq provides very high genome coverage with nucleotide resolution, we were able to obtain RNA abundance and degradation rates at sub-genic resolution. Surprisingly, RNA stability was approximately constant across a transcript. To obtain accurate, position-dependent RNA degradation rates, we had to account for the simultaneously effect of RNA elongation in the data. As a result, we also extracted RNA elongation rates for many genes.

Using the measured parameters, we further analyzed the stoichiometry of transcription and translation machinery throughout the genome. We calculated the RNA synthesis rates and average density of RNA polymerases actively transcribing different genes. By using the protein synthesis rates obtained from the previous chapter, we calculated the density of ribosomes on transcripts. Together, these densities provide a view of the allocation of important gene expression machinery in the cell, and reveal that bacterial cells do not have a strong bias in their strategy for expressing gene products.

As a final analysis, we compared the timescales of the dynamics. The elongation rate gave an average synthesis time per RNA transcript, which together with the RNA degradation rate, can be used to predict co-transcriptional degradation for some RNA transcripts.

While we have developed a new approach to measuring and analyzing RNA dynamics, further technical advances can be made. For example, a pulse-chase protocol will allow us to follow a specific RNA population over time to explore directly the

dynamics of nascent transcripts. We discuss the development of a metabolic labeling technique for RNA in *E. coli* in the next chapter (Chapter 4) that may be useful towards realizing more sophisticated measurements.

3.13 Materials and Methods:

3.13.1 Strains and growth conditions

For the RNA-seq data, an overnight culture of MG1655 cells was diluted 1:250 into LB and grown to OD600nm = 0.3 with shaking at 30C. For qPCR data, AS19 cells were used for the streptolydigin experiment, and corresponding control with rifampicin. The overnight culture was diluted 1:250 in M9 glucose (as above) and grown to OD600nm = 0.15 at 30C.

Rifampicin (Sigma Aldrich) was used at a final concentration of 500ug/ml and streptolydigin (ChemCon GmbH) was used at a final concentration of 100ug/ml.

3.13.2 Spike in RNA

Spike in RNA was prepared from phiX174. The genes D, F, G and H, and fragments 190 (pos 4501 - 4680) and 290 (pos 4031 - 4320) were cloned from ssDNA (New England Biolabs) with the addition of a T7 promoter. RNA was produced by *in vitro* transcription using Ambion's MegaScript kit, according to manufacturer's protocol. RNA integrity was verified by running on a TBE gel (BioRad).

3.13.3 Measuring RNA degradation

Rifampicin or streptolydigin was added to the appropriate cultures at a final concentration of 500ug/ml and 100ug/ml respectively. Aliquots were removed at the intended time points and quenched in a 10% volume of cold phenol:ethanol (9:1). The cells were then harvested by centrifugation and washed once with 0.85% NaCl solution or PBS before storing in -80°C.

3.13.4 Purifying RNA

Frozen cell pellets were resuspended in 1 mg/ml lysozyme TE buffer and lysed by an equal volume of Cell Lysis Buffer (Purgene). Sspike in RNA was added at this point. The suspension was then extracted 1-2 times in 1 volume of acidic phenol/chloroform (OmniPur/Ambion). The aqueous layer was extracted once in chloroform, and RNA was collected by isopropanol precipitation. Contaminating DNA was removed by DNase I (NEB) treatment for 30 min at 37°C, and the resulting RNA was repurified.

3.13.5 Preparing library for Illumina sequencing

To remove ribosomal RNA, we used 1.5ug of total RNA with EpiCenter's Ribo-Zero kit (Bacteria), scaling the reaction to 60% and followed manufacturer's protocol. About 40-80ng of RNA remain.

The purified RNA was fragmented using Ambion's Fragmentation Reagent at 70°C for 5 min, and collected by Zymo's RNA columns. RNA seq libraries were prepared according to Illumna's RNAseq protocol, using NEB enzymes and barcoded adapters (Integrated DNA Technologies). The libraries were sequenced with HiSeq2000

(Center for Systems Biology, Harvard University and Dana-Farber Center for Cancer Computational Biology).

3.13.6 Data processing

For the MG1655 dataset, 49 base reads were mapped onto the MG1655 genome using Bowtie 0.12.7 and then manipulated by homebrewed python 2.6 code. To analyze the data by position, reads were assigned to their chromosome position, and the reads for a 300 nt fragment bin in a transcription unit were counted. The number of reads was normalized by the average of the total number of spike in reads, ssrA, ssrS, and rnpB, before normalizing by abundance at time 0min. The fragments were aligned by the 5' end of the translation start site, and the number of reads was averaged over all transcription units.

Within transcription units, the data was broken into 300 nt bins. The half life for each fragment was obtained by fitting in Igor using the parameters

$$\text{if } 0 < x < P_i\text{: } y = 1$$
$$\text{else, } y = \exp(-(x - P_i)/ T_i)$$

where x is time (min), P_i is the polymerase passage time for fragment i, and T_i is the lifetime of fragment i.

The lifetime of a transcription unit is the weighted average of the lifetime of all fragments.

A weighted linear regression of Pi and the bin position using Matlab was performed to extract elongation rates.

Correlation analyses were performed by Microsoft Excel.

3.13.7 Validation by qPCR

Purified RNA was reverse transcribed according to manufacturer's instructions (MMuLV, New England Biolabs). qPCR reactions were prepared using DyNAmo SYBR green qPCR kit (Finnzymes) and data was collected by 7500 Fast Real-Time PCR system (Applied Biosystems). The cycling parameters followed manufacturer's recommendation. The following primers were used at a final concentration of 0.25uM as probes (listed 5' to 3'):

ssrA1: AAAGAGATCGCGTGGAAGCC
ssrA2: ACCCGCGTCCGAAATTCCTA
rfe1: AACCAAACTTCCGCAAACGTCACC
rfe2: AAAACAAGCA CACCGGCACA AGC
rffH1: ACTCACGACAGCCTGATTGAAGC
rffH2: TTCGCTAATG AACTGGCAGC ACG
rffC1: AGGTGAAGTTGATTTGGCGCTACC
rffC2: CGAAAACGGC TTTGCGCAAA TGC
rffM1: CCAAAGCAGGAGATCATCATGCG
rffM2: AGCGTTTGCC AGATTTTCGG TGC
alaC1.1: TACGCGCATTGATCGTCTCC
alaC1.2: AGTCGCACCGTCCGGGTTAC
alaC2.1: GAAGATGATGATCCTCGGCTTCC
alaC2.2: CACATCGTAGCGTTTCGCCA
alaC3.1: CGAAGCGAAGGTTTGTGTCTCG
alaC3.2: TGGCCTGACGAATACGGTCG

3.14 References:

1. Apirion, D. (1973) "Degradation of RNA in *Escherichia coli* – a Hypothesis" *Mole. Gen. Genet.* **122**, 313 – 322

2. Bai, L., Fulbright, R.M., Wang, M.D. (2007) "Mechanochemical Kinetics of Transcription Elongation" *PRL* **98**, 068103

3. Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S., Cohen, S.N. (2002) "Global Analysis of mRNA Decay and Abundance in *Escherichia coli* at Single-Gene Resolution Using Two-Color Fluorescent DNA Microarrays." *Proc. Nat. Ac. Sci.* **99**, 9697 – 9702

4. Bremer, H., Dennis, P.P. (1995) "*Escherichia coli* and *Salmonella*: Molecular and Cellular Biology" *Editor. Neidhardt F.C.* **Vol 2,** 1553 - 1569

5. Cannistraro, V. J., Kennell, D. (1985) "Evidence that tne 5' end of lac mRNA starts to decay as soon as it is synthesize.d" *J. Bacteriol.* **161**, 820 – 822

6. Carpousis, A.J. (2007) "The RNA Degradosome of *Escherichia coli*: An mRNA-Degrading Machine Assembled on RNase E." *Annu. Rev. Microbiol.* **61**, 71-87

7. Dennis, P.P. Bremer, H. (1973) "Regulation of Ribonucleic Acid Synthesis in *Escherichia coli* B/r: An Analysis of a Shift-up" *J. Mol. Biol.* **75**, 145 -159

8. Epshtein, V., Nudler, E. (2003) "Cooperation between RNA Polymerase Molecules in Transcription Elongation" *Science* **300**, 801 – 804

9. Golding, I., Cox. E.C. (2004) "RNA Dynamics in Live *Escherichia coli* cells." *Proc. Nat. Ac. Sci.* **101**, 11310 – 11315

10. Llopis, P. M., Jackson, A. F., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T., Jacobs-Wagner, C. (2010) "Spatial Organization of the Flow of Genetic Information in Bacteria." *Nature* **466**, 77 – 81

11. Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M. (2007) "Absolute Protein Expression Profiling Estimates the Relative Contributions of Transcriptional and Translational Regulation." *Nature Biotech.* **25**, 117 – 124

12. Miller, O.L. Jr., Beatty, B.R. (1969) "Visualization of Nucleolar Genes" *Science* **164**, 955 – 957

13. Miller, O.L. Jr., Hamkalo, B.A., Thomas, C.A. Jr. (1970) "Visualization of Bacterial Genes in Action" *Science* **169**, 392 – 395

14. Morikawa, N., Imamoto, F. (1969) "On the Degradation of Messenger RNA for the Tryptophan Operon in *Escherichia coli*." *Nature* **223**, 37 – 40

15. Morse, D. E., Mosteller, R., Baker, R. F., Yanofsky, C., (1969) "Direction of *in vivo* Degradation of Tryptophan Messenger RNA – A Correction." *Nature* **223**, 40 – 43

16. Mortazavi, A., Williams, B.A., McCue, K., Schaffer, L., Wold, B. (2008) "Mapping and Quantifying Mammalian Transcriptomes by RNA-seq." *Nature Methods* **5**, 621 – 628

17. Mosteller, R.D., Yanofsky, C. (1970) "Transcription of the Tryptophan Operon in *Escherichia coli*: Rifampicin as an Inhibitor of Initiation" *J. Mol. Biol.* **48**, 525 – 531

18. Roberts, J.W., Shankar, S., Filter, J.J. (2008) “RNA Polymerase Elongation Factors.” *Annu. Rev. Microbiol.* **62**, 211- 233

19. Rose, J.K., Mosteller, R.D., Yanofsky, C. (1970) “Tryptophan Messenger Ribonucleic Acid Elongation Rates and Steady-State Levels of Tryptophan Operon Enzymes Under Various Growth Conditions.” *J. Mol. Biol.* **51,** 541 - 550

20. Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., Rosenow, C. (2003) “Global RNA Half-Life Analysis in *Escherichia coli* Reveals Positional Patterns of Transcript Degradation” *Genome Research* **13**, 216 -223

21. Vogel, U. Sorensen, M., Pedersen, S., Jensen, F., Kilstrup, M. (1992) “Decreasing Transcription Elongation Rate in *Escherichia coli* Exposed to Amino Acid Starvation” *Mol. Microbiol.* **6**, 2191 – 2200

22. Vogel, U., Jensen, K.F. (1994) “The RNA Chain Elongation Rate in *Escherichia coli* Depends on the Growth Rate.” *J. Bacteriol.* **10**, 2807 – 2813

Chapter 4: Metabolic labeling of RNA in *E.coli*

Contributions:

This project was designed and developed by Dr. Katsuyuki Shiroguchi, a postdoctoral fellow, Prof. Sunney Xie, and myself. Dr. Shiroguchi initially screened different nucleotide analogs for compatibility with RNA polymerase and uptake in *E. coli*, while I developed assays to detect incorporation into RNA. After we found a suitable assay, I continued screening nucleotide analogs for uptake, and developed a purification protocol for the labeled RNA.

4.1 Abstract

Metabolic labeling is a general technique used to label recently synthesized molecules in a cell, to answer questions that involve an element of time. Traditionally, labeling has been done using radioactive substrates. The recent development of high throughput sequencing has made it important to consider the use of substrates amendable to purification for labeling nucleic acids. While suitable substrates have been identified for various eukaryotic cells, little attention has been paid to identifying substrates for bacteria. We report here that 4-thiouridine (4sU) can be taken into live *E. coli* cells, and incorporated into RNA. The 4sU-labeled RNA was easily purified and analyzed by techniques like qPCR, and should be compatible with high throughput sequencing.

4.2 Introduction

To construct a physical picture of what is happening inside the cell with our genome-wide RNA elongation and degradation rates, we need to understand how the synthesis and decay of RNA are coordinated. For instance, are most RNA molecules full length, and in existence for a few minutes before they are degraded? Or is synthesis and degradation simultaneous on the same molecule? To answer such questions, we need to look specifically at nascent RNA. We need to isolate nascent RNA through metabolic labeling.

Metabolic labeling of cellular components has traditionally been carried out through the use of radioactive substrates such as ^{14}C-tryptophan, ^{14}C-uracil, and ^{3}H-uracil [7, 11, 12]. It is a general technique that can be used to answer questions that are concerned with time. Because the radioactive label is relatively non-specific, targeting all protein products, or all nucleic acids produced during a defined period of time in the cell, some form of purification is required prior to analyzing the radiolabeled products. This can result in gene-specific information – probes to the trp operon were used to determine the direction of RNA degradation – or general information such as the half-life of proteins [11, 12, 8].

High throughput sequencing naturally processes a population of RNA to give gene-specific information, eliminating the intermediate need for gene-specific purification. If high throughput sequencing can be made compatible with metabolic labeling experiments, such experiments can be made genome wide.

In order to combine the techniques, the label needs to allow the separation of labeled products. Mammalian cells can be incubated in 6-thioguanosine, 4-thiouridine or bromouridine to label nascent RNA [9, 5]. This technique can be extended to yeast if a human transporter is artificially expressed [10]. Efficient purification protocols were developed for these nucleoside analogs [1, 6]. Thus, eukaryotic systems were well-positioned to quickly combine metabolic labeling and high throughput sequencing [2, 10].

Thymine-requiring *Escherichia coli* (*E. coli*) have been made to accept bromouracil to generate labeled DNA [4]. However, there have been no similar efforts to label RNA. Thus to date, nascent RNA experiments in *E. coli* have been carried out with radioactive substrates [3]. There has been one instance where environmental samples of bacteria were incubated with digoxigenin-11-uridine to examine stress response [15]. It was also shown that not all bacterial species respond equally to metabolic labeling [16].

We thus endeavored to identify a suitable nucleoside analog for metabolic labeling of RNA in *E. coli*, for the eventual purpose of studying nascent RNA by high throughput sequencing.

4.3 Identification of a nucleoside analog that labels RNA in live E.coli cells

Besides being able to enter the cell, a suitable nucleoside analog for labeling RNA in live *E. coli* cells needs to be minimally pertubative to RNA generation and degradation to be compatible with intended downstream experiments. In addition, to be useful for purification, this nucleoside analog needs to have a tightly interacting partner that can be used to isolate the labeled RNA. The biotin-streptavidin interaction is one such interacting pair that is commonly used for purification purposes. However, the addition of

a biotin group to a nucleoside may compromise on the molecule's ability to enter the cell and be incorporated into the RNA. A solution is to use a smaller reactive chemical group that can be conjugated with a biotin after RNA is isolated from the cell. Lastly, the nucleoside analog should not be native to the cell so that labeling is specific.

Given the many requirements, we first set out to find suitable substrates that can enter cells and be incorporated into RNA, before evaluating their effect on RNA enzymes and considering other issues.

Because we were pessimistic about the permeability of the bacterial cell wall to small molecules, we initially attempted all experiments in the highly permeable AS19 strain [14]. To check for incorporation of the nucleoside analogs, we developed a dot blot assay based on biotin interaction with a streptavidin-linked fluorophore.

The AS19 cells responded best to 4-thiouridine (4sU) of all the molecules we tried (figure 4.1).

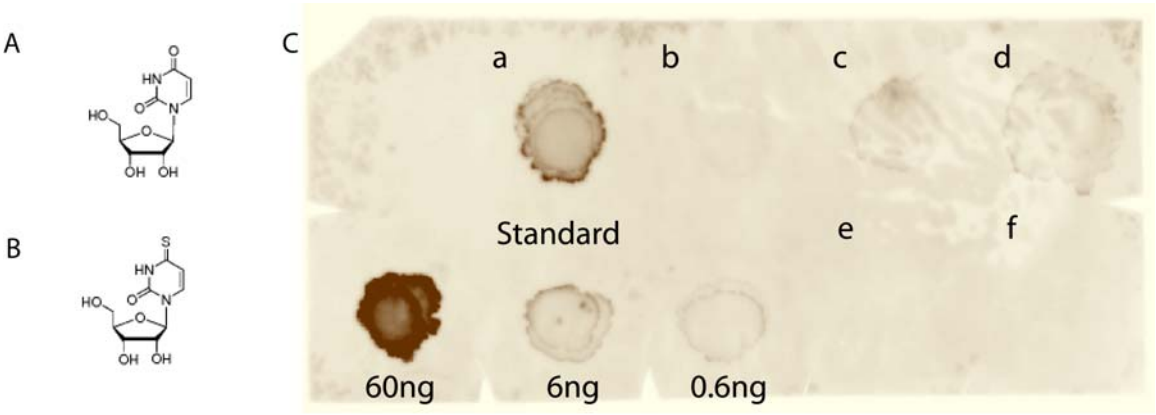


Figure 4.1: AS19 cells readily incorporate 4-thiouridine. (A) Structure of uridine. (B) Structure of 4-thiouridine. (C) A dot blot was prepared with equivalent amounts of purified RNA isolated from cells grown under different conditions. (a, b) Cells incubated with or without 4-thiouridine. (c, d) with or without aminoallyluridine (e, f) with or without ethylyluridine respectively. Signals are compared to that of known amounts of *in vitro* synthesized RNA labeled with biotin-UTP.

Surprisingly, wildtype MG1655 cells were better at incorporating 4sU than AS19 cells (figure 4.2). This may be due to the fact that MG1655 cells are healthier and grow more quickly than AS19 cells.

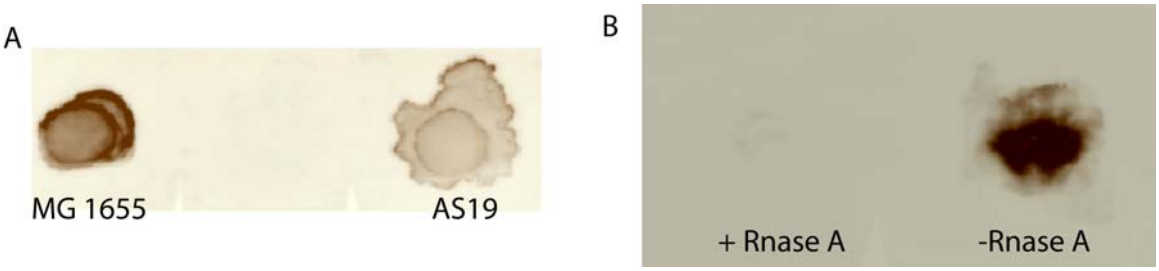


Figure 4.2: MG1655 cells incorporate 4-thiouridine more readily into their RNA than AS19 cells. (A) MG1655 and AS19 cells were grown up in the presence of 4sU for the same amount of time. The same amounts of purified RNA were used for the dot blot. MG1655 cells incorporate 1.5x more 4-thiouridine than AS19. (B) Incubating the labeled RNA with RNase A destroys the signal.

4.4 At low concentrations, 4-thiouridine does not perturb RNA elongation and degradation

High concentrations of 4sU were used to check for incorporation in live cells. However, it was obvious that cells were not healthy under those conditions. We titrated the concentration of 4sU down, and accordingly saw growth rates increase for cells grown in M9 glucose (table 4.1).

[4-thiouridine]	Change in OD_{600nm} after 1 hour	
	M9 glucose	LB
4.8mM	0.04	--
2.4mM	0.07	0.27
1.2mM	0.08	0.28
0.6mM	0.10	0.31
0.2mM	--	0.35
0mM	0.12	0.29

Table 4.1: 4-thiouridine at high concentrations slow cell growth in M9 glucose. MG1655 cells grown in M9 glucose or L Broth (LB) media were exposed to different amounts of 4-thiouridine. Cell growth was measured by $OD_{600nm.}$ Initial OD_{600nm} was 0.09 for all M9 glucose samples, and 0.14 for all LB samples.

We found that 4sU incorporation into RNA saturates quickly, such that increasing the concentration of 4sU in the media does not result in increased labeling. We also checked the effect of 4sU on RNA elongation and degradation on a few genes by quantitative polymerase chain reaction (qPCR). We noticed that the rplN gene seemed a little more sensitive to 4sU concentration than the other genes we looked at (cyoA, cyoE, and rnb), which generally had a less than 2-fold change in RNA half life (Figure 4.3). Using the lifetime of rplN as an indicator, we decided to use 0.2mM of 4sU for cells grown in M9 glucose

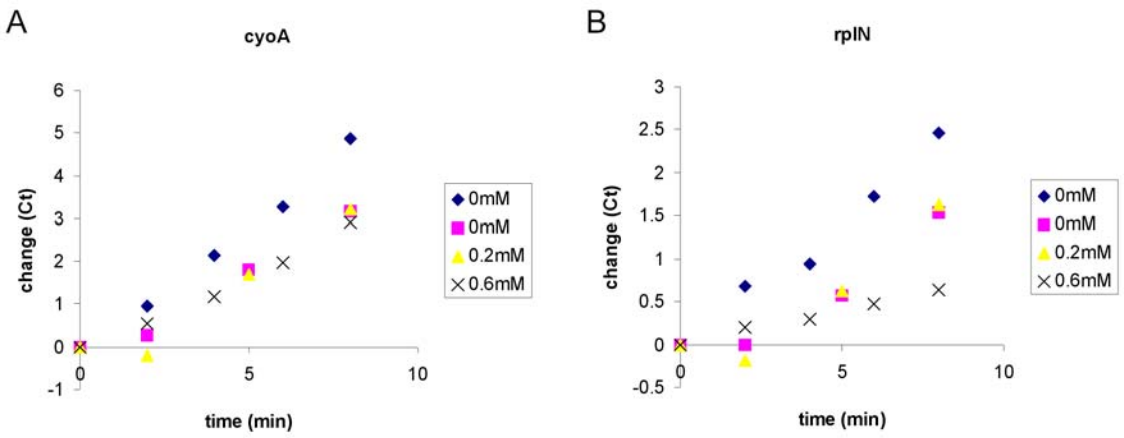


Figure 4.3: Lifetime of rplN RNA seems sensitive to 4-thiouridine. Data is combined from two separate experiments. Set 1: time points at 0, 2, 4, 6, 8 minutes for 0mM 4sU sample(blue/diamond) and 0.6mM 4sU(cross). Set 2: time points at 0, 2, 5, 8 minutes for 0mM (pink/square) and 0.2mM (yellow/triangle). (A) the cyoA has lifetime of 1.7-2.4 min at 0mM 4sU. The lifetime becomes 2.3min at 0.2mM 4sU and 2.8min at 0.6mM 4sU. (B) rplN RNA has a lifetime of 3.4-5.1 min at 0mM 4sU. The lifetime becomes 4.6min at 0.2mM 4sU and 12.9min at 0.6mM 4sU.

We tried growing cells in L broth to obtain higher cell densities. MG1655 cells were less sensitive to 4sU when grown in LB (table 5.3), but also incorporated less 4sU into RNA. Cells tolerate up to 1.2mM of 4sU before the halflife of rplN changes.

4.5 Purification and quantification of labeled rRNA

We chose to label our 4sU-RNA with iodoacetyl-PEG2-biotin, instead of the chemically detachable HPDP-biotin, and use streptavidin-coated magnetic beads to purify labeled RNA, slightly modifying existing protocols [10, 1, 17]. Following successful purification of *in vitro* synthesized RNA (Figure 4.4), we proceeded with ribosomal RNA (rRNA) as our next target. rRNA is the most abundant RNA species in the cell, and should be easily purified if our system works.

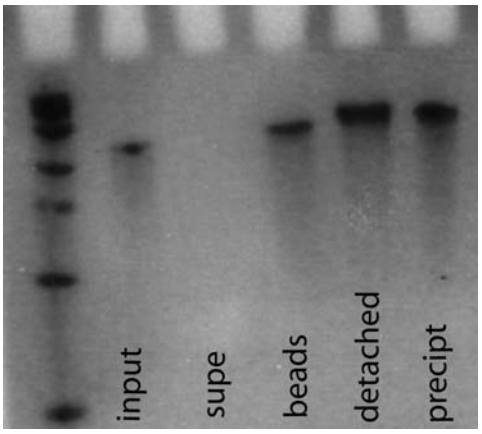


Figure 4.4: Magnetic bead protocol successfully purifies intact labeled RNA. An *in vitro* synthesized, biotin-labeled ~400nt test RNA sample was followed through the various steps of the purification protocol. Similar amounts of the RNA were expected in each well of a 5% acrylamide gel. An RNA ladder is included in the first lane. The supernatant (supe) was taken from the first wash, and the beads were collected after the last wash. The buffer in which the RNA was detached from the beds is loaded (detached). Precipitated RNA was resuspended and loaded (precipt).

We collected total RNA from cells incubated in 4sU for different amounts of time, and purified the labeled RNA. After reverse transcription, we probed for 23S rRNA using qPCR (Figure 4.5). Although there was a ~30-fold increase in qPCR signal after 60 minutes of incubation, the actual difference in the amount of RNA purified was only 2-fold. This suggests that most of the RNA purified from the 0 min sample may be contaminated by non-specific binding or some natively thiolated RNA species.

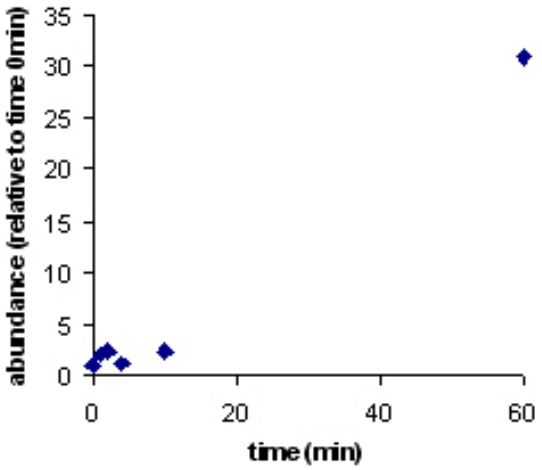


Figure 4.5: Successful purification of labeled 23S rRNA from MG1655 cells. Cells were incubated in 0.2mM 4sU (M9 glucose) and collected at different times (0,1, 2, 4,10, 60 min). The amount of labeled 23S rRNA is quantified by qPCR and normalized by labeled spiked in RNA, which does not account for cell division during the incubation time.

While native thiol-containing RNA has not been reported to be a problem in mammalian cells, as much as 70% of tRNA in *E. coli* may have some type of thiol group [13]. If thiol-containing tRNA is saturating the magnetic beads (in the case that bead capacity is limited), a simple size exclusion column that removes RNA smaller than 200 nucleotides should improve the purification of labeled 23S rRNA. Indeed, the amount of labeled RNA increases by 600-fold after 40 min if the sample if passed through a column that removes 75% of RNA shorter than 200nts (Figure 4.6).

4.6 Kinetics of 4-thiouridine incorporation into RNA

At the current state of technology, any application of 4sU metabolic labeling in *E. coli* that requires resolution on the order of tens of minutes is feasible. However, the processes we are interested in studying, namely RNA degradation and elongation, require a time resolution of a minute or less. We thus need to characterize the kinetics of 4sU incorporation into RNA on that time scale.

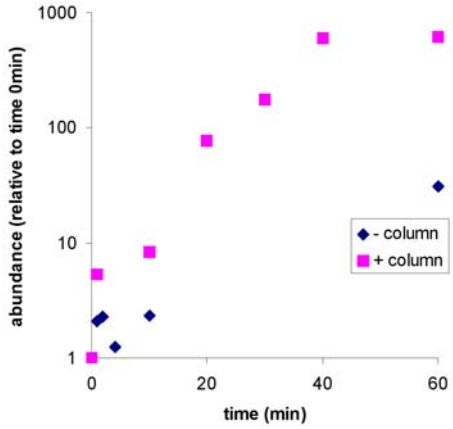


Figure 4.6: Purification efficiency of 4sU-labeled 23S rRNA is improved using a column. Cells were incubated in 0.2 mM 4sU (M9 glucose) and collected after different incubation times (0, 10, 20, 30, 40, 50, 60 min). The RNA was subjected to an additional column purification step before bead purification (pink/square). The result of the previous purification (without column) is plotted (blue/diamond) for comparison. This reveals that the amount of labeled 23S rRNA was previously underestimated because the beads were saturated by native thiol-containing RNA.

After many failed attempts to detect the incorporation of 4sU into RNA on a minute timescale by qPCR, we re-visited the use of dot blots. We were able to easily measure the labeled RNA to estimate that there is a lag of about 3- 4 minutes before labeled RNA is produced by the cell (Figure 4.7).

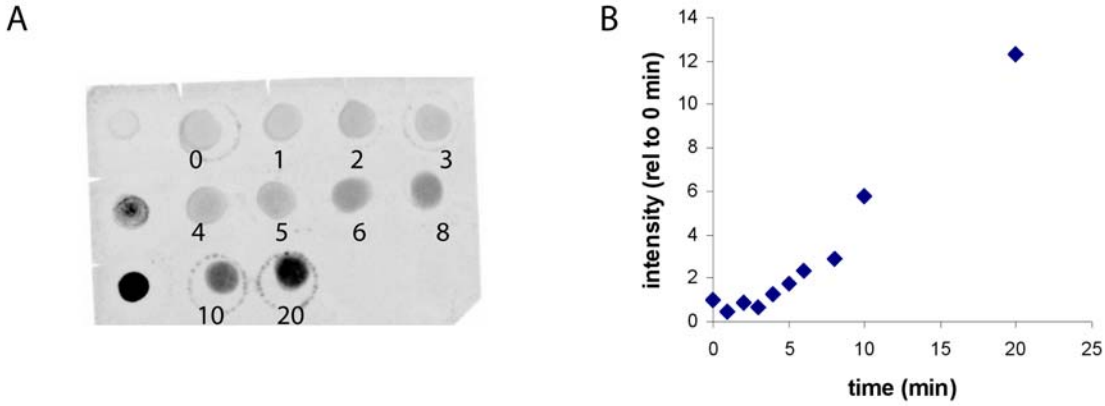


Figure 4.7: A short lag in 4sU incorporation into RNA. Cells were grown in LB with 1.2mM 4sU. (A) Total RNA was spotted on a membrane. The numbers correspond to the time in 4sU in minutes. The 3 spots on the left correspond to 0.6ng, 6ng and 60ng of *in vitro* synthesized RNA. (B) The intensities from the dots were quantified and plotted, revealing that there is a lag in 4sU incorporation into RNA.

4.7 Summary and Outlook

We have identified a nucleoside analog, 4sU, that is useful for the metabolic labeling of RNA in *E. coli* cells. 4sU penetrates the cell and gets incorporated into RNA within minutes. It is reactive with iodoacetate, allowing the attachment of a biotin group that permits easy and efficient purification of labeled RNA. Although thiol-containing RNA molecules are native to *E. coli*, we are able to remove most of the contaminating species through size selection by a column.

Currently, the 4sU labeling method works well on the order of tens of minutes, similar to 4sU labeling in mammalian cells [10, 1]. This allows for gene profiling in *E. coli* in response to perturbation. A potentially interesting experiment to try is to label *E. coli* persisters with 4sU to see if they are indeed as metabolically dormant as expected (see appendix for discussion of bacterial persisters).

With further characterization of the lag in 4sU incorporation into RNA and, if necessary, experimenting with different conditions, we may be able to push the time resolution of the method down to minutes, comparable to the time scale of RNA degradation and synthesis in bacteria.

Although we have been focused on 4sU and RNA, the feasibility of metabolic labeling in *E. coli* suggests that similar nucleoside analogues may be used to study newly synthesized DNA.

4.8 Materials and methods

4.8.1 Strains and growth conditions

MG1655, a K-12 sub-strain, and AS19, a mutant *E. coli* B strain with a leaky membrane, were used. Strains were grown at 30C in either M9 medium with 0.4% glucose, or L Broth (Miller). Typically, an overnight culture is inoculated at 1:200 for M9 glucose or 1:250 for LB until OD_{600nm} is 0.1 or 0.3 respectively (approximately 3 to 4 hours). MG1655 is sensitive to contaminated M9 glucose media and will grow poorly.

We tried various nucleoside analogs for labeling RNA, including biotin-cytosine, biotin-uridine, bromouridine, ethylyluridine and aminoallyluridine. 4-thiouridine (Sigma Aldrich) was diluted to 96mM, and the stock solution was stored in -20C.

For RNA half life measurements, cells were incubated in 4sU for 1 hour before addition of rifampicin and harvest, as described previously (see 3.14).

4.8.2 Purification of labeled RNA

RNA was purified as previously described (see 3.13). A MegaClear column (Applied Biosystems) was used to remove small RNAs, including tRNAs, in some experiments. If necessary, the RNA is fragmented in 0.5x fragmentation reagent (Applied Biosystems) at 70C for 3 minutes.

A biotin handle was added using EZ-link iodoacetyl-PEG2-biotin (Pierce, subdivision of ThermoFisher). The iodoacetyl-PEG2-biotin was dissolved in water at 10mg/ml just before the reaction. Up to 100ug of RNA was added to 10ul of iodoacetyl-PEG2-biotin solution, 7.5ul of 10x PBS (Lonza), 1ul of 0.5M EDTA and the reaction was made up to 75ul with water. The reaction was incubated at room temperature (~25C) in the dark for 2 – 4 hours.

The RNA was purified by ethanol precipitation, and resuspended in TE buffer. Unlike the HPDP-biotin reaction, no chloroform extraction step is needed to remove excess biotin.

40ul of streptavidin magnetic beads (New England Biolabs) per 50ug of total RNA were pre-washed in low salt buffer (0.15M NaCl, 20mM Tris HCl (pH7.5), 1mM EDTA), then incubated with yeast tRNA (Sigma Aldrich) in low salt buffer for at least 15 min at room temperature with rotation. The beads were then resuspended in 150ul high salt buffer (0.5M NaCl, 20mM Tris HCl (pH7.5) and up to 50ul of RNA is added. The RNA and beads were incubated for 30 min at RT with rotation. The beads were then washed 2x with 100ul high salt buffer, 2x with low salt buffer, 2x with 4M urea (4M urea, 10mM Tris-HCl (pH7.5), 1mM EDTA) (16) and finally once more in low salt buffer. The beads were resuspended in 20ul biotin solution (2.5mM biotin, Tris-base to pH7.4, 1mM EDTA) (16). 20ul of 95% formamide (Applied Biosystems) and 1ul of 0.5M EDTA (pH8.0) were added to the biotin RNA solution, and the mixture was heated at 95C for 5 min to release the RNA.

After removing the beads, the supernatant was extracted once in phenol/chloroform (pH4.5, Sigma Aldrich) and precipitated.

4.8.3 In vitro synthesized RNA

To generate a control and reference, phiX174's D gene was cloned as a template for *in vitro* transcription. RNA was synthesized using the MegaScript kit (Applied Biosystems) with biotin-UTP added to 1/5 of the UTP concentration. The RNA was checked by polyacrylamide gel (BioRad) and stored in -80C. We used the 0.1-1kb RNA ladder from USB(Affymetrix).

The control phiX174 D RNA was used as a standard in dot blots, and as a spike in in qPCR experiments.

4.8.4 RNA Dot Blot

Intact total RNA was labeled with biotin. Zeta-probe blotting membrane (Biorad) was pre-washed in 20x SSC buffer for at least 1 hour at RT and dried. A denaturation buffer (50ul 95% formamide, 16.2ul 37% formaldehyde, 10ul 10x MOPS buffer (200mM MOPS, 50mM sodium acetate (pH7.0), 10mM EDTA)) was added to total RNA at a ratio of 3:1 volume. The RNA was heated at 65C for 5 min then placed immediately on ice. 1 volume of cold 20x SSC buffer was added to the RNA mixture (making it a final of 3:1:4 denaturation buffer, RNA, 20x SSC).

The RNA mixture was spotted onto the membrane in 2ul drops and allowed to dry. For large volumes, it is helpful to use a vacuum setup to draw the RNA into the membrane. The membrane was placed onto wet blotting paper (use 10x SSC) and exposed to 1200mJ UV (UV crosslinker, VWR) to crosslink the RNA.

The membrane was pre-incubated in 50mM BSA (New England Biolabs) in PBS+0.1% Tween20+5mM EDTA (PBSTE) for 30 min at RT, then streptavidin-Alexa488 (diluted to 2ug/ml in TBS) (Invitrogen) to PBSTE+BSA at 1:1000 was added. The membrane was incubated for 30 min RT then washed 3x 10 min in PBSTE. The membrane was imaged with a Typhoon Imager (Amersham), using the Alexa488 filter and 488nm excitation laser.

4.8.5 qPCR

The lifetime experiments were performed as previously described (see 3.13).

For the quantification of labeled 23S rRNA, total RNA was fragmented before it was biotinylated. Purified labeled RNA was reverse transcribed according to manufacturer's instructions (MMuLV, New England Biolabs). qPCR reactions were prepared using DyNAmo SYBR green qPCR kit (Finnzymes) and data was collected by 7500 Fast Real-Time PCR system (Applied Biosystems). The cycling parameters followed manufacturer's recommendation. The following primers were used at a final concentration of 0.25uM as probes (5' to 3'):

23S rRNA1: AAGACCAAGGGTTCCTGTCCAACG
23S rRNA2: TACACGCTTAAACCGGGACAACC
phiX174 D1: TTGGATTGCTACTGACCGCTCTCG
phiX174 D2: ACAGGCCGTTTGAATGTTGACGG

4.9: References

1. Cleary, M.D., Meiering, C.D., Jan, E., Guymon, R., Boothroyd, J.C. (2005) "Biosynthetic Labeling of RNA with Uracil Phosphoribosyltransferase Allows Specific Microarray Analysis of mRNA Synthesis and Decay." *Nature Biotech* **23**, 232 – 237

2. Core, L.J., Waterfall, J.J., Lis, J.T. (2008) "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters." *Science* **322**, 1845-1848

3. Epshtein, V., Nudler, E. (2003) "Cooperation Between RNA Polymerase Molecules in Transcription Elongation" *Science* **300**, 801-805

4. Hewitt, R., Suit, J.C., Billen, D. (1967) "Utilization of 5-bromouracil by Thymineless Bacteria" *J. Bact.* **93**, 86-89

5. Jackson, D.A., Iborra, F.J., Manders, E.M.M., Cook, P.R. (1998) "Numbers and Organization of RNA polymerases, Nascent Transcripts, and Transcription Units in HeLa Nuclei" *Mol. Biol. Cell* **9**, 1523-1536

6. Kenzelmann, M. et al. (2007) “Microarray Analysis of Newly Synthesized RNA in Cells and Animals” *Proc. Nat. Acad. Soc.* **104**, 6164 – 6169

7. Kjedlgaard, N.O. (1961) “The Kinetics of Ribonucleic Acid and Protein Formation in Salmonella typhimurium During the Transition Between Different States of Balanced Growth” *Biochim. Biophys. Acta* **49**, 64-76

8. Koch, A.L., Levy, H. R. (1955) “Protein Turnover in Growing Cultures of *Escherichia coli*” *J. Biol. Chem.* **217**, 947-958

9. Melvin, W.T., Milne, H.B., Slater, A.A., Allen, H.J., Keir, H.M. (1978) “Incorporation of 6-thioguanosine and 4-thiouridine into RA. Application to Isolation of Newly Synthesized RNA by Affinity Chromatography” *Eur. J. Biochem.* **92**, 373-379
10. Miller, C. et al. (2011) “Dynamic Transcriptome Analysis Measures Rates of mRNA Synthesis and Decay in Yeast.” *Mol. Sys Biol.* **7**, 458

11. Morikawa, N., Imamoto, F. (1969) “On the Degradation of Messenger RNA for the Tryptophan Operon in *Escherichia coli*” *Nature* **223**, 37-40

12. Morse, D. E., Mosteller, R., Baker, R. F., Yanofsky, C. (1969) “Direction of *in vivo* Degradation of Tryptophan Messenger RNA – A Correction” *Nature* **223**, 40-43

13. Prasadarao, Y.S., Cherayil, J.D. (1972) “Separation of ^{35}S-labeled Thionucleosides of *E. coli* and *P. aeruginosa* tRNAs on a Phosphocellulose Column” *Biochim. Biophys. Acta* **299**, 1-7

14. Sekiguchi, M., Iida, S. (1967) “Mutants of *Escherichia coli* permeable to actinomycin” *Proc. Nat. Ac. Sc.* **58**, 2315-2320

15. Stankiewicz, N., Gold, A., Yuksel, Y., Berensmeier, S., Schwartz, T. (2009) “*In vivo* labeling and specific magnetic bead separation of RNA for Biofilm Characterization and Stress-induced Gene Expression Analysis in Bacteria” *J. Microbiol. Meth.* **79,** 344-352

16. Urbach, E., Vergin, K.L., Giovannovi, S.J. (1999) “Immunochemical Detection and Isolation of DNA from Metabolically Active Bacteria” *App. Env. Microbiol.* **65**, 1207-1213

17. Wilson, D. (1998) “Using Immobilized Streptavidin” Szostak lab website, http://genetics.mgh.harvard.edu/szostakweb/protocols/biotin-avidin/index.html

Appendix 1: Uncovering genetic interactions in a bacterial persistence network through transposon sequencing

Contributions:

Prof. Sunney Xie, Prof. Eric Ruben (Harvard School of Public Health), and I designed the initial experiments for studying persisters. I performed preliminary experiments to characterize the emergence of persister cells over time. Later, Dr. Katsuyuki Shiroguchi, a postdoctoral fellow, and I further conceived the application of transposon sequencing to persisters. I constructed the necessary plasmids and strains, and performed the mating to generate transposon libraries. .Dr. Guoqing Zhang, a former postdoctoral fellow, assisted with transposon library construction and collection of persister cells. Dr. Haifeng Duan and Dr. Xiaohui Ni, former and current postdoctoral fellows, respectively, assisted in preparation of the transposon-sequencing (tn-seq) library. Dr. Ni and Prof. Hao Ge (Peking University) assisted with data analysis.

A1.1 Abstract

When an isogenic population of bacteria is treated with antibiotics, not all cells are killed at the same rate. The slow dying cells are called persisters. Many genes have been identified to cause persistence. To construct a network of persistence genes, we performed genetic interaction studies in *Escherichia coli* using saturating transposon libraries to screen for interaction partners of hipA and hipA7, the mutant hipA allele linked to the high persistence phenotype. Genes exhibit very different fitness in the two backgrounds. Of the 256 putative interaction partners identified to interact with either

allele, only one gene is expected to interact with both hipA and hipA7. Many interaction partners are from disparate pathways, or of unknown function, making it hard to piece together a network. This may support the hypothesis that many diverse pathways are capable of generating persistence.

A1.2 Introduction

Persistence is the phenomenon where an isogenic population of cells exhibits two different dying rates when incubated homogenously in antibiotics. The first and main population dies very rapidly, but a very small fraction dies at a much slower rate. Since the 1980's, there have been many efforts in *Escherichia coli* to identify key genes responsible for causing bacterial persistence [11, 12, 17, 18, 21, 0]. While several proteins have been identified as potentially important players, no single gene has been proven to cause persistence. In addition, if all the candidate genes are truly related to persistence, it is not clear whether persistence is the result of these different candidate proteins acting in multiple pathways, or in the same pathway.

To date, the most famous persistence-related protein is hipA. A mutated version of the protein, hipA7, is known to increase persistence by 100-1000 fold [18]. hipA7 has been used in multiple studies to understand *E.coli* persisters [2, 8, 13, 14]. The mechanism of hipA7 action is unknown, but hipA is structurally related to kinases [5], and has been found to interact with EF-Tu [20].

While researchers have focused their efforts on hipA and hipA7, they have not considered the pathway that hipA functions in. By expanding our understanding of the pathway that hipA or hipA7 works in, we may be able to integrate the other persistence-

related genes that have been identified to date, to create a deeper understanding of persistence. A quick and efficient method to map out pathways is through the use of genetic interaction studies.

Genetic interaction studies identify genes that function in the same or similar pathways, creating a map of pathways and networks in the cell. A strain is constructed with two deleted genes, and depending on the growth and viability of this strain with respect to the single deletion strains, we can obtain information about the relationship between the two genes. Through the use of mating or conjugation, large numbers of double-deletion strains were created for yeast *Sacchromyces cerevisiae* and bacteria *E. coli*, permitting large scale genetic interaction studies [3, 22, 23].

Another method of generating genetic interaction data is by transposon. The gene of interest is deleted in a strain of bacteria, and transposon insertions randomly inactivate a second DNA locus. The efficiency of transposition limits the number of insertions in each cell to one. In a saturating transposon library, where many cells with transposon insertions are used, we interrogate all possible interacting genes. The transposon is engineered to allow easy identification of the insertion site through a genome-wide method like microarray, or high throughput sequencing [1, 7, 19, 24]. This method allows a one pot experiment, simplifying the study.

In this section, we describe preliminary attempts to study the genetic interactions of hipA and hipA7 through the use of a transposon library.

A1.3 Construction of three transposon libraries

We adopted the transposon scheme designed by van Opijnen et al. [24] Three transposon libraries were constructed by mating– library 1 started with the wildtype MG1655 strain to provide reference fitness values for hipA, library 2 started with a MG1655/hipA7 strain to provide reference fitness values for hipA7, and library 3 started with MG1655 with deleted hipA, which also serves as a hipA7 deleted strain – each with at least 100,000 colonies to ensure that all genes are adequately sampled.

In a typical experiment to uncover genetic interaction, the growth or fitness of the doubly disrupted strain is compared to that of the single disruption strains. In our experiment, we compare the ability of the doubly disrupted strains to generate persisters to that of singly disrupted strains. The cells are treated with antibiotics until the concentration of surviving cells, measured by colony forming units (cfu), stabilizes. In a single strain experiment, we would directly compare cfu across different strains. In this one-pot transposon experiment, we sequence the insertions to find which doubly-disrupted strains are represented differently than expected in the population.

A1.4 Persisters arise from entry into stationary phase

We initially found that there were almost no persisters in our transposon libraries (data not shown). During the construction of the transposon libraries, cell proliferation was minimized to avoid introducing unintended selection pressures to the libraries. This meant that the cells were never allowed to enter stationary phase, which is proposed to be important in the generation of persisters. For the experiment to proceed, we needed to allow persisters to form.

It is important to identify the exact instance when persisters are generated to minimize unintended selection pressures in our experiment. There is slight disagreement

in literature about when persisters are generated during stationary phase – an initial report by Keren et al. [13] proposed that persisters arise from entry into stationary phase, while a later report by Gefen et al. [8] proposed that persisters arise an hour after exit from stationary phase. Part of the discrepancy may be attributed to the subtle differences in the method of introducing antibiotics to the samples. We thus re-investigated the relationship between stationary phase and persisters.

In their experiments, Keren et al. [13] directly added antibiotics to an aliquot of cells. Antibiotics are known to be most effective when used against actively growing cells, and less effective in cells that are in stationary phase. Gefen et al. [8] diluted their aliquot of cells into fresh media with antibiotics. This allows cells from stationary phase to grow up and be more effectively killed. We thus adopted the protocol from Gefen et al. [8].

We measured the largest increase in the number of persisters upon entry into stationary phase (figure A1.1). The number of persisters increases about ~1000-fold in hipA7 strain and ~300-fold for wildtype MG1655 cells. In comparison, Gefen et al. [8] measured a 10-fold increase in the number of persisters at exit from stationary phase. We noticed that their measurement did not extend into the entry of stationary phase. Our numbers are expectedly smaller than those measured by Keren et al. [13], which showed a 10^4-fold increase in surviving cells for both the wildtype and hipA7 strains.

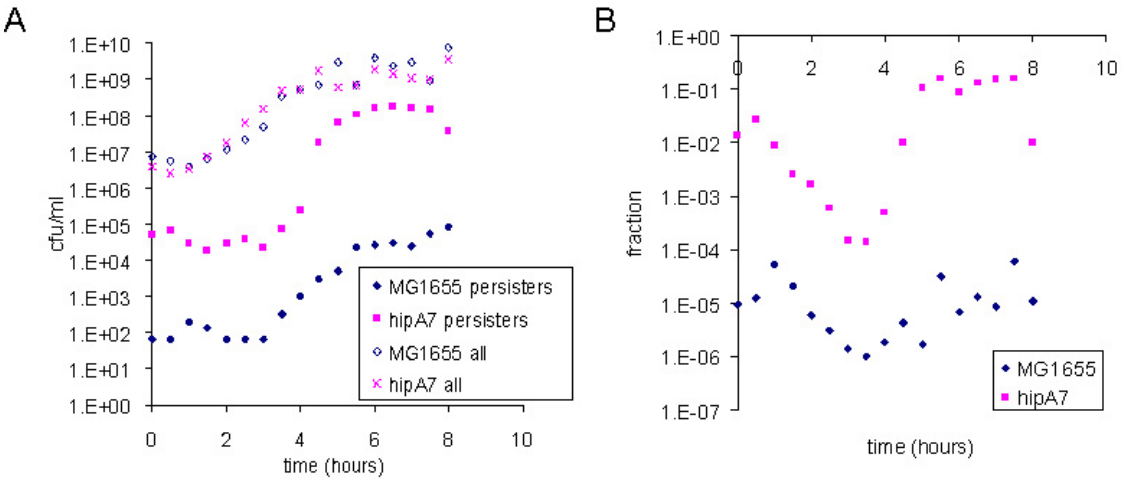


Figure A1.1 Persisters are formed during entry into stationary phase for wildtype MG1655 strain and a hipA7 MG1655 strain. (A) Absolute cfu/ml counts (y axis) over time are plotted for all cells (empty diamond, MG1655 or crosses, hipA7 strain) or persisters (diamond, MG1655, or squares, hipA7 strain) for a representative experiment. (B) The fraction of persisters relative to the population for MG1655 (diamond) or hipA7 strain (square). The number of persisters may rise slightly initially after exit from stationary phase, but the biggest rise in persister numbers occurs at entry into stationary phase (~1000 fold for hipA7). The ΔhipA strain behaves like the wildtype MG1655 strain.

Accordingly, the transposon libraries were grown for 6 hours before the addition of antibiotics to allow persisters to form. Persisters were allowed to grow up after 3.5 hours of carbenicillin treatment and 1,000,000 colonies were collected per library. Genomic DNA was purified from these samples, and suitable sequencing libraries were prepared according to previously published protocol [24].

A1.5 Calculating the fitness of genes

We calculated the fitness of each insertion for each transposon library, and averaged the fitness score over a gene, modifying the analysis by van Opijnen et al. [24]. We confirmed that the redundant IS150 sequences had a fitness of about 1 (hipA: 1.01± 0.02; hipA7: 1.03 ± 0.08; ΔhipA: 1.00 ± 0.02) and are a suitable reference for genes that have no fitness effect. We normalized all fitness numbers using the IS50 fitness numbers. The mean fitness in all three strains is about 1, but the spread of fitness is larger for the

hipA7 strain (figure A1.2). We note that the gene fitness varies more widely in the hipA7 strain. The fitness of hipA is 0.99, while the fitness of hipA7 is 1.24. These numbers are consistent with the observation that hipA7 is important for the cell to survive antibiotic exposure.

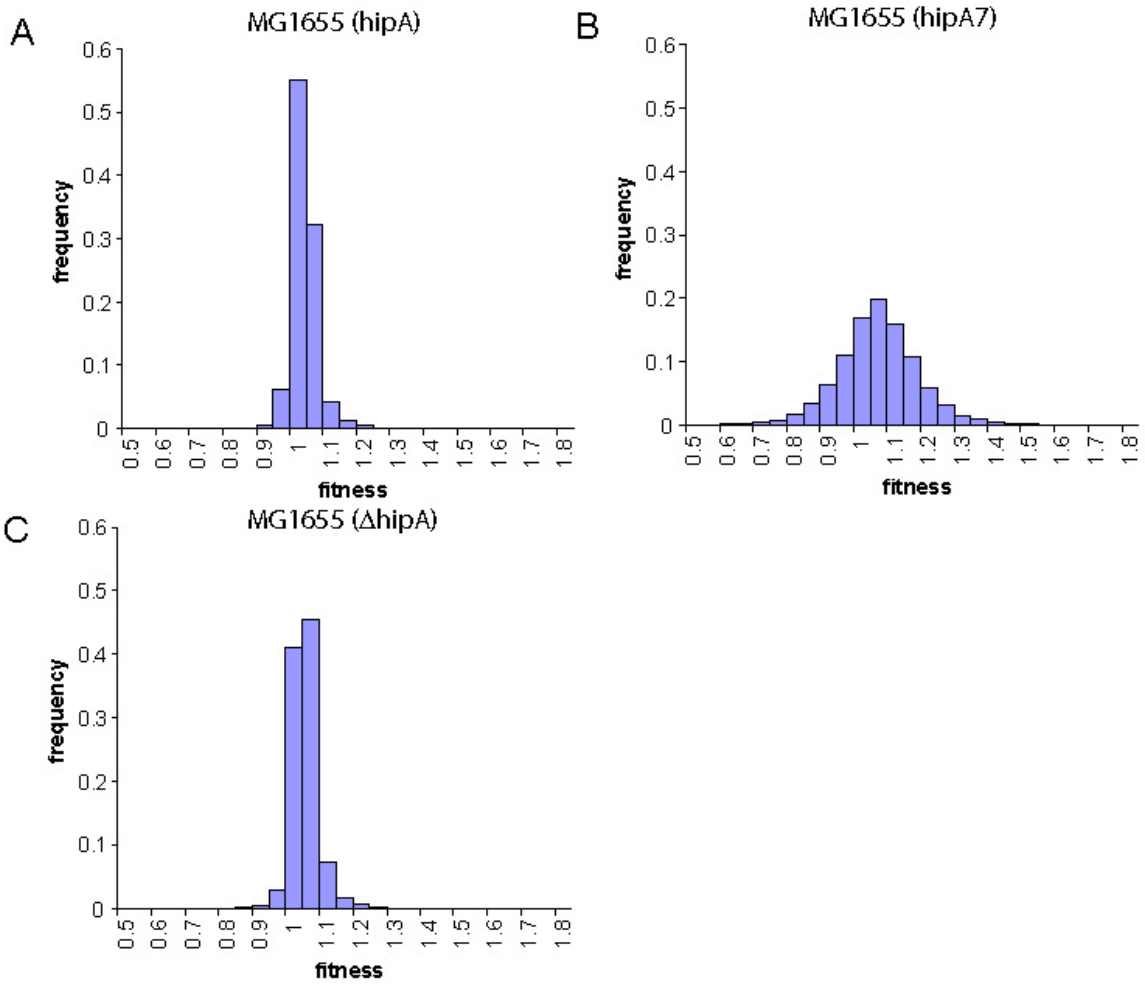


Figure A1.2 Fitness of genes in different strains (A) The fitness of 3296 genes in wildtype MG1655(hipA) are plotted. The mean fitness is 1.01 ± 0.04 (standard deviation). (B) The mean fitness of 3184 genes in the MG1655(hipA7) strain is 1.05 ± 0.12. (C) The mean fitness of 3290 genes in MG1655 (ΔhipA) is 1.01 ± 0.04.

We confirmed that most of the previously annotated essential proteins had a fitness of 0 (hipA: 293/302; hipA7: 299/302; ΔhipA: 297/302) (PEC database). For the

essential proteins that had a fitness score, some of the results can be explained by the modular structure of proteins. The fitness of some proteins was domain-dependent such that the protein can sustain insertions in the non-essential domains and have a fitness score (figure A1.3). This was also observed in a normal growth type transposon experiment with *Mycobacteria smegmatis* (Jason Zhang, Rubin lab, personal communication).

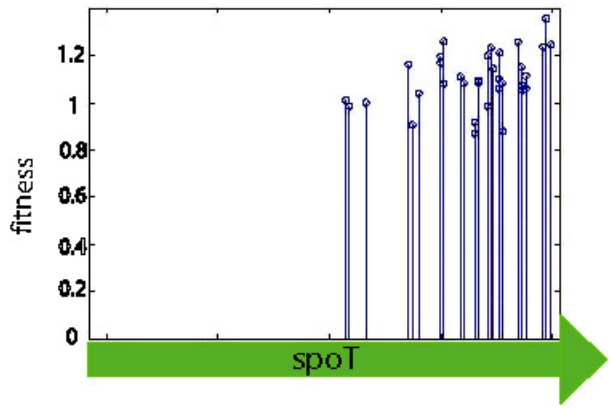


Figure A1.3 Essential and non-essential domains in spoT, a protein that has been classified as essential. Each position where a transposon insertion was detected is marked, and the fitness of the insertion is plotted. We only observed insertions in the C terminal half of the gene. *E. coli* spoT is an essential gene that has hydrolase and ppGpp synthetase activity in the N terminal half of the protein [9].

A1.6 Correlation between strains

While it is known that the number of persisters in strains with hipA and hipA7 are very different, it is not clear to what extent the strains behave differently. We compared the fitness of genes in each of the three strains (figure A1.4). The best correlation is observed between the hipA and the ΔhipA strains. The hipA7 strain correlates relatively poorly with the hipA and ΔhipA strains. Although the hipA and hipA7 alleles only differ

by two point mutations, these differences effect a profound change in the way the cell is wired.

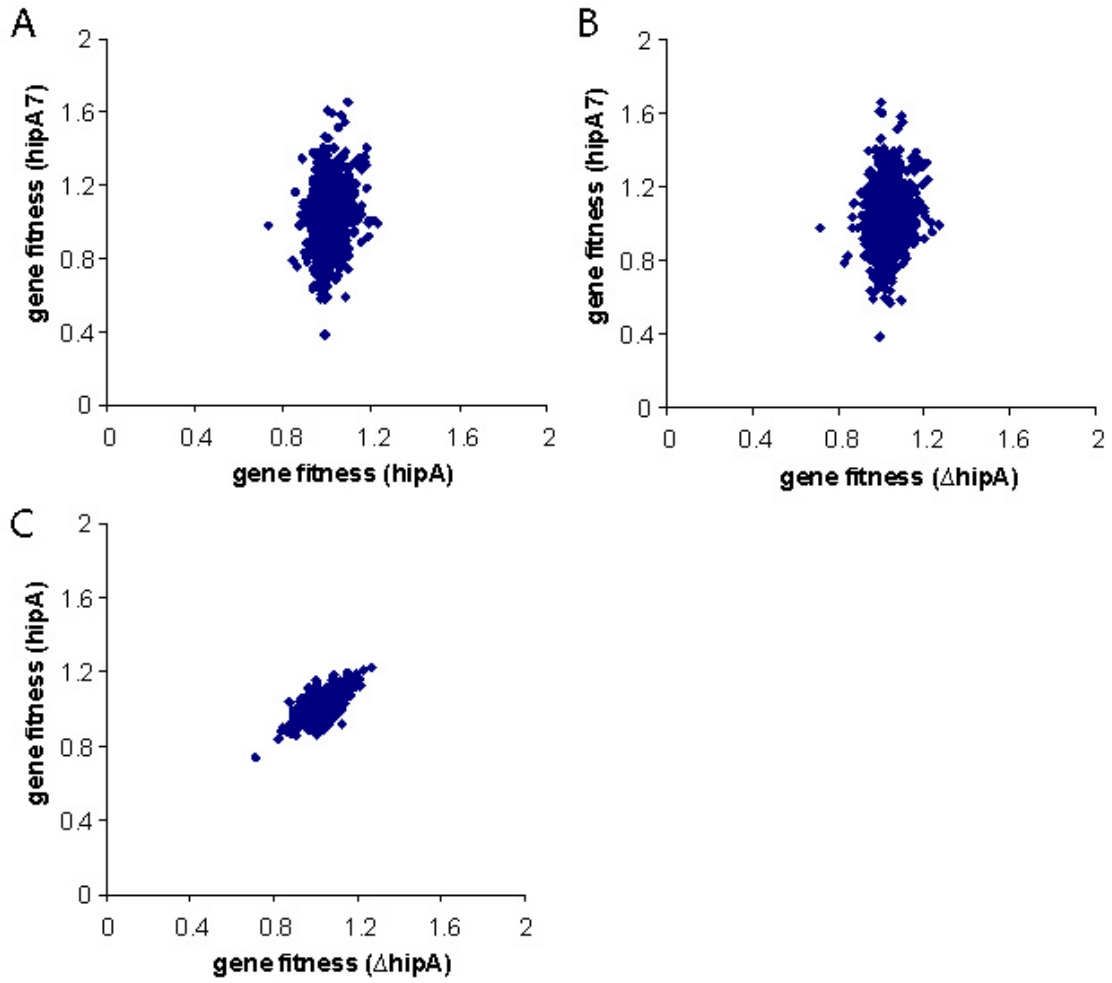


Figure A1.4 Correlation of gene fitness across three strains (A) Correlation between hipA and hipA7 strains is 0.14. (B) The correlation between ΔhipA and hipA7 is 0.16. (C) The correlation between ΔhipA and hipA is 0.57.

A1.7 Potential interaction partners of hipA and hipA7 alleles

To obtain potential interaction partners of the hipA and hipA7 genes, we compared the differences between the double mutant (observed value of ΔhipA strain) and the fitness of the single disruptions (expected value) for genes that had at least 5

insertions in each dataset. van Opijnen et al. [24] used a 10% threshold to determine interactions. Most genes (3086/3167) deviate less than 10% from the expected fitness, and thus can be considered to have no interaction with the hipA gene (figure A1.5). On the other hand, about a third of genes (1037/2944) deviate from the expected fitness by more than 10% with the hipA7 allele. This is likely an effect of the wide variance in gene fitness in the hipA7 strain. To limit the set of genes for the current discussion, we set the threshold at where the deviation sharply increases with the queried genes (deviation from a moving average of 100). This is 11% for hipA dataset (35 genes), and 40% for the hipA7 dataset (27 genes).

hipA and hipA7 have no interaction partners in common (see Supplementary table A1.6 for full list of genes). The lack of overlap between the two sets of genes suggests that hipA7 is a gain of function mutation that operates in very different pathways in the cell compared to the original hipA protein.

A third of identified interaction partners are genes of unknown function according to *E. coli* database EcoCyc (9/30 for hipA7 and 10/28 for hipA). The genes of known function are from disparate pathways. We also identified ryjA, a small RNA of unknown function, as an interaction partner of hipA. The lack of congruency in the genes potentially involved in the persistence network makes it difficult to propose a mechanism for hipA/hipA7 action. However, the diversity of proteins involve suggests that persistence could result from many pathways.

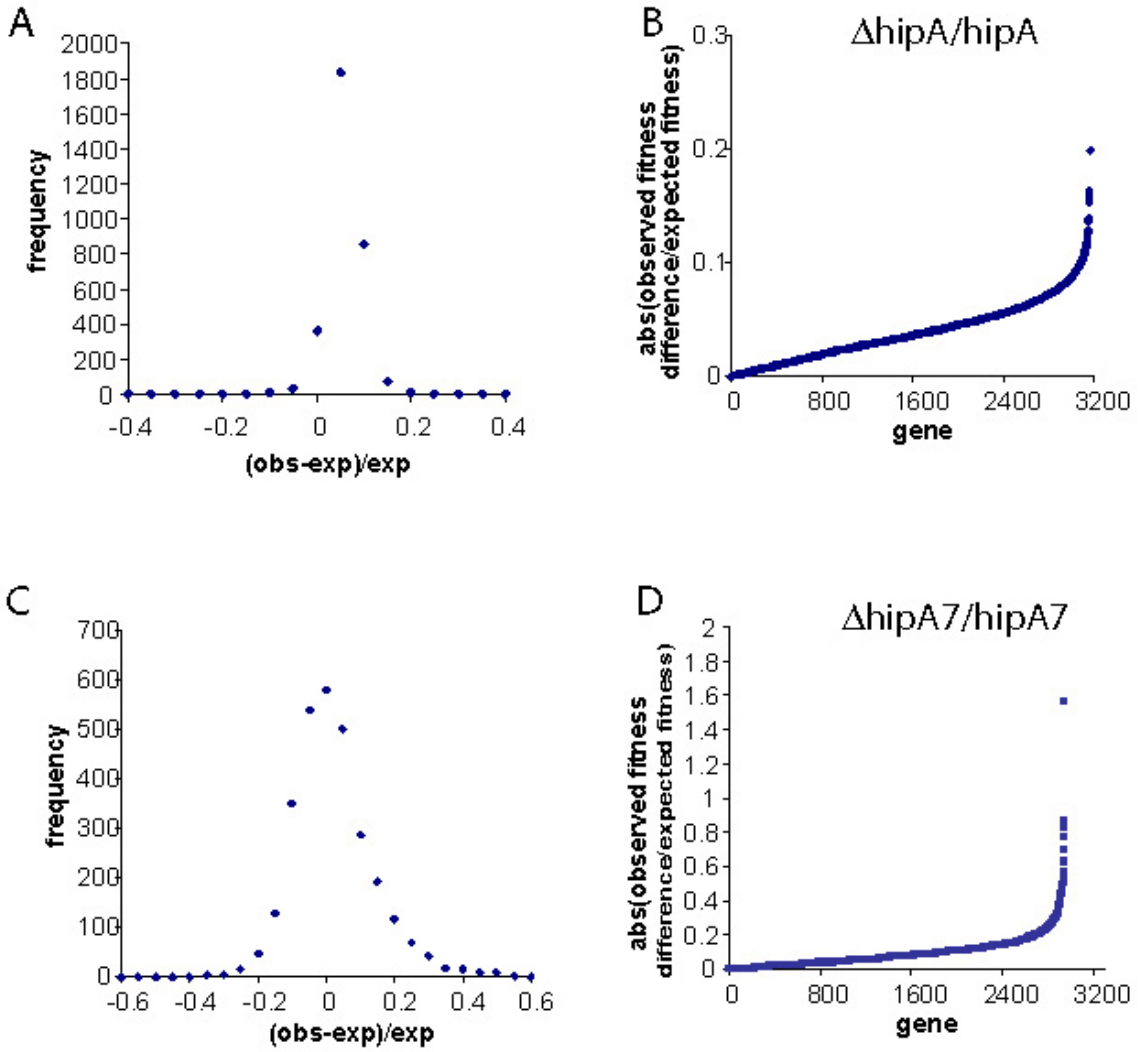


Figure A1.5 Comparing observed fitness of double disruption to the expected fitness of single disruptions (A) The observed fitness for the hipA dataset (3167 genes, which had at least 5 insertions in each dataset) were narrowly distributed at 0.03+0.03 (mean + std dev). (B) The absolute difference between the observed and expected fitness (scaled as a fraction of the expected value) is typically small in the hipA strain, and increases sharply past 10%. (C) The observed fitness for the hipA7 dataset (2944 genes) were distributed at 0.00 + 0.13 (mean + std dev). (D) The absolute difference between the observed and expected fitness (scaled as a fraction of the expected value) is typically large in the hipA7 strain, and increases sharply past 30%. There is a relatively larger difference in the observed and expected fitness for hipA7.

A1.8 Synthetic lethal interactions

In the previous section, genes with zero insertions in either library were removed from analysis. This includes genes with synthetic lethal interactions, which will have insertions in the hipA or hipA7 background, but no insertions in the ΔhipA strain. To ensure that the fitness in the hipA/hipA7 is robust, we set a higher threshold of at least 10 insertions per gene. This results in two potential synthetic lethal genes in the hipA strain, yshB and hycB, which did not appear in the list of potential interaction partners of hipA7.

For hipA7, 149 genes are potentially synthetic lethal. All 149 genes do not appear to interact with hipA. The full list of potential synthetic lethal genes is reported (Supplementary table A1.7).

The list of interactions will need to be verified by experiment. It is puzzling why genes related to flagella synthesis may have a fitness effect in the generation of persisters in an *E. coli* strain not known to be particularly mobile.

A1.9 Suppressor interactions

Suppressor interactions have a pattern of insertions that is the opposite of synthetic lethal interactions: the single disruption in the hipA/hipA7 strain has no reads, but when hipA is deleted, the doubly disrupted strain is viable. As in the case of the synthetic lethal interactions, we set a higher threshold of 10 insertions per gene in the doubly disrupted strain. We add the extra caveat that the hipA7 samples have slightly lower coverage than the other two samples, making the likelihood of a false positive higher.

44 suppressor interactions were observed for hipA7 and 3 for hipA (Supplementary table A1.8). ykgL, which has synergistic interactions with hipA, is a suppressor of hipA7.

A1.10 Conclusion and outlook

We have constructed three saturating transposon libraries to identify genes that interact with the persistence-related proteins, hipA and hipA7, with the goal of understanding the network of genes that cause persistence. We observed that the fitness of genes in a hipA7 strain fluctuates greatly compared to a hipA or ΔhipA strain. In addition, the fitness of genes are more similar in the hipA and ΔhipA strains compared to in the hipA7 background.

We identified a total of 256 candidate genes that interact in the same or similar pathways as persistence-related proteins, hipA and hipA7. While some of these interactions have subtle effects (insertions can be detected in the gene in both the single or doubly disrupted strains), some interactions are crucial to generating persisters (no insertions detected in the gene in either the single or doubly disrupted strains). Of the 256 genes, only ygkL, a gene of unknown function, is proposed to interact with both hipA and hipA7. It interacts synergistically with hipA, but suppresses hipA7. Given the many differences between the hipA and hipA7 strains, it may be important to distinguish between both alleles in future experiments.

We have used a slightly different definition of fitness in our experiment than in previously published studies. The fact that many genes in the hipA7 strain have fitness that are different from 1 raises the question of whether the fitness criterion used is

unsuitable, or whether genes in the hipA7 strain have very different contributions in the cell – hipA7 could genuinely have more interactions than hipA, and the hipA7 cell is wired very differently compared to wildtype MG1655. To distinguish between the two situations, fitness should be measured during regular exponential growth for the hipA and hipA7 strains. In a growth experiment, genes in a hipA7 background may not display such a wide range of fitness.

Beyond the specific questions about persisters, the results of this transposon screen suggest that two point mutations may be sufficient to drastically change the contributions of genes in a cell.

A1.11 Materials and methods

A1.11.1 Plasmid and strain construction

The K-12 sub-strain, MG1655, was the basis for all three transposon libraries. The hipA deletion strain was constructed by replacing the hipA gene with a chloramphenicol gene by λ-Red recombination [6, 26](Yu 2000, Datsenko 2000). The hipA7 strain, TH1269, was obtained from Thomas Hill [15]. All strains were transduced with P1 phage to insert a rpsL150 mutation conferring streptomycin resistance.

The inverted repeat sequence on plasmid pSC189 [4] was mutated to insert an MmeI restriction site. The ampicillin resistance was switched to chloramphenicol by isothermal *in vitro* recombination [10]. The plasmid, now called pSC189m7cam, is propagated in SM10 λpir cells.

A1.11.2 Transposon library construction

SM10λpir/pSC189m7cam cells were mated with the MG1655/rpsL150(hipA, hipA7 or ΔhipA) for 3 hours at 37C. The cells were plated onto LB plates with 35mg/ml Kan and 50 mg/ml streptomycin and allowed to grow overnight at 37C. At least 100,000 colonies were collected for each library and stored in 15% glycerol at -80C.

A1.11.3 Time of persister formation

Overnight cultures of MG1655 and TH1269 were diluted 1:1000 into pre-warmed LB-peptone broth (where the tryptone is replaced by peptone) and grown at 37C with shaking. 200ul of the culture was diluted into 1.8ml of pre-warmed fresh LB-peptone with 50mg/ml carbenicillin and incubated for 4 hours at 37C. The carbenicillin was inactivated by addition of penicillinase (BD Biosciences), and appropriate dilutions were plated onto LB plates. The number of colonies was counted the next day.

A1.11.4 Isolating persisters from transposon library

Frozen cells were inoculated into pre-warmed LB-peptone at 1:1000, and grown up for 6.5 hours at 37C with shaking. 1ml of culture was reserved and stored. This is the before drug sample. The culture was diluted 1:10 into pre-warmed LB-peptone with 50mg/ml carbenicillin and incubated for another 3.5 hours at 37C. The culture was treated with penicillinase, and diluted 1:2 with 30% glycerol and stored in -80C.

1,000,000 colonies were grown up on LB with kanamycin and streptomycin for each library. They were collected and stored at -80C. This is the after drug sample.

A1.11.5 Preparation of sequencing libraries

For each transposon library, there is a before drug and an after drug sample. Genomic DNA was purified from each sample using the Puregene kit (Gentra).

The DNA was digested with MmeI (New England Biolabs), and the sequencing libraries were constructed according to previously published protocol [24]. We used Phusion polymerase (Finnzymes) at an annealing temperature of 61C for 20-22 cycles to amplify our library.

Six samples were run in one lane of HiSeq2000 with 1x50bp reads (FAS Center for Systems Biology, Harvard University). The summary statistics are reported in supplementary table A1.9.

A1.11.6 Data analysis

The 16 base reads were mapped by Bowtie to the MG1655 genome (NC000913, NCBI) allowing 0 mismatches [24]. Fitness scores were calculated for insertions with at least 5 reads using the following equation [16, 24]

$$W_i = \frac{\ln(N_i(t_2)\times d / N_i(t_1))}{\ln((1-N_i(t_2))\times d/(1-N_i(t_1)))}$$

in which $N_i(t_1)$ and $N_i(t_2)$ are the frequency of the mutant in the population at the beginning and the end of the experiment, and d, which is typically is the expansion factor, representing the growth of the bacterial population, is the fraction of persisters. This is 10^{-4} for the hipA and ΔhipA strains, and 0.05 for hipA7.

For genes that had at least 5 insertions, we averaged the fitness of the insertions to calculate the fitness of the genes. Finally we normalized the fitness to with the fitness of the ins genes, which are expected to have no fitness effect ($W = 1$).

A1.12 References

1. Badarinarayana, V., Estep III, P.W., Shedure, J., Edwards, J., Tavazoie, S., Lam, F., Church, G.M. (2001) "Selection Analyses of Insertional Mutants using Subgenic-resolution Arrays." *Nat. Biotech.* **19**, 1060 – 1065

2. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., Leibler, S., *Science* 305 : 1622-1625, (2004)

3. Butland, G., Babu, M. et al. (2008) "eSGA: E. coli synthetic genetic array analysis" *Nat. Methods* **5**, 789 – 795

4. Chiang, S.L., Rubin, E.J. (2002) "Construction of a *mariner*-based transposon for epitope-tagging and genomic targeting." *Gene* **296**, 179 – 185

5. Correia, F.F., D'Onofrio, A., Rejtar, T., Li, L., Karger, B.L., Makarova, K., Koonin, E.V., Lewis, K. (2006) "Kinase Activity of Overexpressed HipA is Required for Growth Arrest and Multidrug Tolerance in *Escherichia coli.*" *J. Bacteriol.* **188**, 8360 – 8367

6. Datsenko, K.A., Wanner, B.L. (2000) "One-step Inactivation of Chromosomal Genes in *Escherichia coli* K-12 Using PCR Products." *Proc. Nat. Ac. Sci.* **97**, 6640 – 6645

7. Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D.V., Akerley, B.J. (2009) "Tracking Insertion Mutants Within Libraries by Deep Sequencing and a Genome-wide Screen for *Haemophilus* Genes Required in the Lung." *Proc. Nat. Ac. Sci.* **106**, 16422 – 16427

8. Gefen, O., Gabay, C., Mumcuoglu, M., Engel, G., Balaban, N. Q., *PNAS* 105: 6145 – 6149, (2008)

9. Gentry, D.R., Cashel, M. (1996) "Mutational analysis of the Escherichia coli spot gene identifies distinct but overlapping regions involved in ppGpp synthesis and degradation." *Mol. Microbio.* **19**, 1373 – 1384

10. Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison III, C.A., Smith, H.O. (2009) “Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases.” *Nat. Methods* **6**, 343 – 345

11. Hansen, S., Lewis, K., Vulic, M., *Antimicrob. Agents Chemother.* 52: 2718 – 2726, (2008)

12. Hu, Y., Coates, A. R. M., *FEMS Microb. Letters* 243 : 117-124 (2005)

13. Keren, I., Kaldalu, N., Spoering, A., Wang, Y., Lewis, K., *FEMS Microb. Letters* 230 : 13-18, (2004)

14. Keren, I., Shah, D., Spoering, A., Kaldalu, N., Lewis, K., *J. Bacteriology* 186: 8172-8180, (2004)

15. Korch, S.B., Henderson, T.A., Hill, T.M. (2003) “Characterization of the *hipA7* Allele of *Escherichia coli* and Evidence that High Peristsence is Governed by (p)ppGpp Synthesis.” *Mol. Microbiol.* **50**, 1199 – 1213

16. Lenski, R.E., Rose, M.R.S., Tadler, S.C. (1991) “Long-term Experimental Evolution in Escherichia coli. I Adaptation and divergence during 2000 generations.” *Am. Nat.* **138**, 1315 – 1341

17. Li, Y, Zhang, Y., *Antimicrob. Agents Chemother.* 51:2092 – 2099, (2007)

18. Moyed, H. S., Bertrand, K. P., *J. Bacteriology* 155: 768 – 775, (1983)

19. Sassetti, C.M., Boyd, D.H., Rubin, E.J. (2001) “Comprehensive Identification of Conditionally Essential Genes in Mycobacteria.” *Proc. Nat. Ac. Sci.* **98**, 12712 – 12717

20. Schumacher, M.A., Piro, K.M., Xu, W., Hansen, S., Lewis, K., Brennan, R.G. (2009) “Molecular Mechanisms of HipA-Mediated Multidrug Tolerance and Its Neutralization by HipB.” *Science*, **323**, 396 – 401

21. Spoering, A. L., Vulic, M., Lewis, K., *J. Bacteriology* 188:5136-5144, (2006)

22. Tong, A.H. et al. (2001) “Systematic genetic analysis with ordered arrays of yeast deletion mutants.” *Science* **294**, 2364 – 2368

23. Typas, A. et al. (2008) “High-throughput, quantitative analyses of genetic interactions in E. coli.” *Nat. Methods* **5**, 781 – 787

24. van Opijnen, T., Bodi, K.L., Camilli, A. (2009) “Tn-seq: High-throughput Parallel Sequencing for Fitness and Genetic Interaction Studies in Microorganisms.” *Nat. Methods* **6**, 767 – 772

25. Wolfson, J. S., Hooper, D. C., McHugh, G. L., Bozza, M. A., Swartz, M. N., *Antimicrob. Agents Chemother.* 34: 1938 – 1943, (1990)

26. Yu, D., Ellis, H.M., Lee, E.-C., Jenkins, N.A., Copeland, N.G., Court, D.L. (2000) "An Efficient Recombination System for Chromosome Engineering in *Escherichia coli*." *Proc. Nat. Ac. Sci.* **97**, 5978 – 5983

hipA interactions		hipA7 interactions	
gene	**deviation**	**gene**	**deviation**
arsR	0.13	citX	0.50
artQ	0.16	dapF	0.44
fimZ	0.15	dps	0.43
glcG	0.12	hha	0.48
hybE	0.14	hycI	0.48
mpaA	0.13	hyfC	0.48
nohQ	0.16	hyfH	0.46
nrdH	0.13	lamb	0.50
nudJ	0.14	malt	0.53
ompL	0.13	mlaB	0.69
ompR	0.14	pdxK	0.44
rpiB	0.12	qseB	0.41
ryjA	0.12	rlmB	0.83
uidR	0.12	rutB	0.63
usg	0.14	scpB	0.64
yadK	0.12	speG	0.87
yafP	0.13	tehB	0.44
yahC	0.14	yafO	0.53
ybfA	0.16	ycjD	0.48
ybiA	0.13	ycjZ	0.57
ybiI	0.12	ydfE	0.47
yciS	0.12	yecD	1.57
yejM	0.12	yeeX	0.49
yfgD	0.16	ygcR	0.77
ygiZ	0.20	ygdR	0.41
yjhC	0.13	yhaJ	0.44
ykgH	0.16	yncB	0.47
ykgL	0.14		
yphF	0.13		
yzgL	0.16		

Supplementary Table A1.1 List of potential interaction partners of hipA or hipA7 35 genes had a fitness that deviated more than 11% from expected for hipA, while 27 genes has a fitness that deviated more than 40% for hipA7. Positive numbers indicate that disruption of the gene and hipA/hipA7 resulted in fewer persisters than expected (syngergy).

Supplementary Table A1.2 List of synthetic lethal interactions with hipA or hipA7

hipA synthetic lethal interactions/# insertions		hipA7 synthetic lethal interactions/# insertions					
hycB	15	aaaE	13	paaF	11	yedQ	48
yshB	11	acnB	11	paaG	22	yedS	25
		afuC	29	paaH	25	yeeD	10
		amyA	31	paaJ	13	yeeR	21
		arnF	23	paaK	18	yeeV	11
		casA	28	paaZ	20	yehB	37
		casB	11	pcnB	16	yfbK	26
		casC	22	peaD	12	yfbP	17
		casD	32	perR	29	yfcO	42
		casE	10	pgm	23	ycfP	14
		dam	29	quuQ	10	yfcQ	10
		dcm	28	relE	11	yfcR	10
		envy	10	rfaQ	16	yfdM	10
		exoD	18	Mla	54	yfdS	14
		fliA	16	stfP	11	yffL	21
		fliC	45	stfQ	19	yffO	12
		fliD	16	sufA	13	yffP	13
		fliE	10	tfaD	16	yfjK	47
		fliH	12	wbbK	14	yfjL	37
		fliL	12	wbbL	17	yfjM	10
		fliM	12	xapA	20	yfjP	24
		fliN	11	xapB	28	yfjQ	28
		fliY	19	xapR	12	yfjT	12
		flu	36	yabP	18	yfjV	20
		fucP	22	yafW	19	yfjZ	11
		gale	14	yafX	14	ygaQ	16
		galK	22	yafY	23	ygbF	19
		galT	13	yafZ	23	ygbT	18
		gtrA	13	yagA	22	ygcB	68
		ihfA	10	yagB	17	yhcD	109
		intA	31	yagE	13	yhcE	17
		intE	22	yagF	12	yhdF	37
		intF	52	yagG	13	yjaA	11
		intS	49	yagH	27	yjfZ	14
		intZ	16	yagJ	32	yjhF	23
		jayE	21	yagK	12	yjhG	49
		lsrA	22	yagN	10	yjhH	11
		lsrC	24	yaiP	27	yjhI	10
		lsrK	15	ybcH	14	yjiR	28
		lsrR	22	ybcY	28	yjiT	47
		mmuM	22	ybdO	10	yiiI	35
		mmuP	33	ybhA	13	ykfA	31
		modA	13	ybhI	22	ykfB	11
		modC	31	ydbA	74	ykfC	30
		mode	14	ydfR	10	ymfQ	12
		modF	15	ydfV	10	ynaA	14
		nfrA	63	yecS	15	yoaA	25
		nfrB	46	yedA	15	ypjA	89
		ompT	15	yedE	31	yqeI	11
		paaA	10	yedK	11		

Supplementary Table A1.2 (Continued) List of synthetic lethal interactions with hipA or hipA7 The number of insertions for each gene in the hipA/hipA7 strain is listed. Zero insertions were detected for the gene in the doubly disrupted strain. The minimum number of insertions used for the list is 10. 149 genes were identified for hipA7, while only 2 were identified for hipA.

hipA suppressor interactions/# insertions		hipA7 suppressor interactions/# insertions			
arcA	12	abgA	45	rimP	16
pspD	14	abgB	65	sgcB	18
yigG	12	abgR	22	uspE	16
		abgT	99	yahL	11
		atpI	18	ybeL	15
		caiF	15	ybjQ	10
		cedA	14	ydaL	14
		dbpA	28	ydaM	31
		dmsD	12	ydaN	43
		ebgC	14	ydaQ	15
		fnr	31	ydiQ	19
		fucU	11	ydiT	11
		ibsE	10	yfbT	21
		insP	20	ygiV	23
		lon	79	yhcB	12
		mazE	10	yigP	12
		nudG	13	yjfJ	21
		nuoI	15	yjhV	11
		nuoN	22	ykgL	11
		phnF	11	ymcE	19
		phnO	10	yobD	13
		recA	12	yrfF	87

Supplementary Table A1.3 List of suppressor interactions with hipA or hipA7 The number of insertions for each gene in the doubly disrupted strain is listed. Zero insertions were found for the gene in the hipA/hipA7 strains. The minimum number of insertions used for the list is 10. 44 genes were identified for hipA7, while only 3 were identified for hipA.

sample	Total reads	Total insertions (>5 reads/insertion)	# genes with >10 insertions	Average read/insertion
hipA (before)	9,878,947	243,247 (184,074)	3253	40
hipA (after)	9,506,998	205,494 (160,084)	3075	46
hipA7 (before)	13,826,123	186,676 (144,953)	3255	74
hipA7 (after)	5,906,106	124,693 (83,731)	2547	47
ΔhipA (before)	8,720,308	155,941 (145,204)	3115	55
ΔhipA(after)	1,045,594	193,818 (152,956)	3059	53

Supplementary Table A1.4 Summary statistics of sequencing run The data is from one lane on the Illumina HiSeq2000.